SORT 31 (1) January-June 2007, 3-44

ISSN: 1696-2281
www.idescat.net/sort

Objective Bayesian point and region estimation in location-scale models

José M. Bernardo
Universitat de València, Spain

Abstract

Point and region estimation may both be described as specific *decision problems*. In point estimation, the action space is the set of possible values of the quantity on interest; in region estimation, the action space is the set of its possible credible regions. Foundations dictate that the solution to these decision problems must depend on both the utility function and the prior distribution. Estimators intended for general use should surely be invariant under one-to-one transformations, and this requires the use of an invariant loss function; moreover, an objective solution requires the use of a prior which does not introduce subjective elements. The combined use of an invariant information-theory based loss function, the *intrinsic discrepancy*, and an objective prior, the *reference prior*, produces a general solution to both point and region estimation problems. In this paper, estimation of the two parameters of univariate location-scale models is considered in detail from this point of view, with special attention to the normal model. The solutions found are compared with a range of conventional solutions.

MSC: Primary: 62F15, 62C10; secondary: 62F10, 62F15, 62B10

Keywords: Confidence Intervals, Credible Regions, Decision Theory, Intrinsic Discrepancy, Intrinsic Loss, Location-Scale Models, Noninformative Prior, Reference Analysis, Region Estimation, Point Estimation.

1 Introduction

Point and region estimation of the parameters of location-scale models have a long, fascinating history which is far from settled. Indeed, the list of contributors to the simpler examples of this class of problems, estimation of the normal mean and estimation of the normal variance, reads like a *Who's Who* in 20th century statistics.

Address for correspondence: José M. Bernardo is Professor of Statistics at the Universitat de València. Departamento de Estadística e I. O., Facultad de Matemàticas, 46100-Burjassot, Valencia, Spain.
jose.m.bernardo@uv.es, www.uv.es/bernardo
Received: March 2006

In this paper, an objective Bayesian decision-theoretic solution to both point and region estimation of the parameters of location-scale models is presented, with special attention devoted to the normal model. In marked contrast with most approaches, the solutions found are *invariant* under one-to-one reparametrization.

1.1 Notation

Probability distributions are described through their probability density functions, and no notational distinction is made between a random quantity and the particular values that it may take. Bold italic roman fonts are used for observable random vectors (typically data) and bold italic greek fonts for unobservable random vectors (typically parameters); lower case is used for variables and upper case calligraphic for their dominion sets. The standard mathematical convention of referring to functions, say $f_x(\cdot)$ and $g_x(\cdot)$ of $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, respectively by $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ is often used. Thus, the conditional probability density of observable data $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ given $\boldsymbol{\omega}$ is represented by either $p_x(\cdot \,|\, \boldsymbol{\omega})$ or $p(\boldsymbol{x} \,|\, \boldsymbol{\omega})$, with $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) \geq 0$, $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and $\int_{\boldsymbol{\mathcal{X}}} p(\boldsymbol{x} \,|\, \boldsymbol{\omega})\, d\boldsymbol{x} = 1$, and the posterior density of a non-observable parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ given data $\boldsymbol{x}$ is represented by either $\pi_{\boldsymbol{\theta}}(\cdot \,|\, \boldsymbol{x})$ or $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$, with $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \geq 0$ and $\int_{\boldsymbol{\Theta}} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = 1$. Density functions of specific distributions are denoted by appropriate names. In particular, if x has a normal distribution with mean μ and standard deviation σ, its probability density function will be denoted $\mathrm{N}(x \,|\, \mu, \sigma)$, and if λ has a gamma distribution with parameters α and β, its probability density function will be denoted $\mathrm{Ga}(\lambda \,|\, \alpha, \beta)$, with $\mathrm{E}[\lambda] = \alpha/\beta$, and $\mathrm{Var}[\lambda] = \alpha/\beta^2$.

It is assumed that available data $\boldsymbol{x}$ consist of one observation from the family $\mathcal{F} \equiv \{p(\boldsymbol{x} \,|\, \boldsymbol{\omega}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\omega} \in \Omega\}$ of probability distributions for $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$, and that one is interested in point and region estimation of some function $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\omega}) \in \boldsymbol{\Theta}$ of the unknown parameter vector $\boldsymbol{\omega}$. Often, but not necessarily, data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ of some simpler model $\{q(x \,|\, \boldsymbol{\omega}),\, x \in \mathcal{X},\, \boldsymbol{\omega} \in \Omega\}$, in which case, $\boldsymbol{\mathcal{X}} = \mathcal{X}^n$ and the likelihood function is $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) = \prod_{j=1}^{n} q(x_j \,|\, \boldsymbol{\omega})$. Without loss of generality, the original parametric family $\mathcal{F}$ may be written as

$$\mathcal{F} \equiv \{p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\, \boldsymbol{x} \in \boldsymbol{\mathcal{X}},\, \boldsymbol{\theta} \in \boldsymbol{\Theta},\, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}\} \tag{1}$$

in terms of the vector of interest $\boldsymbol{\theta}$, and a vector $\boldsymbol{\lambda}$ of nuisance parameters. A *point estimator* of $\boldsymbol{\theta}$ is some function of the data $\tilde{\boldsymbol{\theta}}(\boldsymbol{x}) \in \boldsymbol{\Theta}$ such that, for each possible set of observed data $\boldsymbol{x}$, $\tilde{\boldsymbol{\theta}}(\boldsymbol{x})$ could be regarded as an appropriate proxy for the actual, unknown value of $\boldsymbol{\theta}$. A *p-credible region* of $\boldsymbol{\theta}$ is some subset $C_p(\boldsymbol{x}, \boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$ whose posterior probability is p. Within this framework, attention in this paper focuses on problems where data consist of a random sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from a *location-scale* model

$m(x\,|\,\mu,\sigma,f)$, of the form

$$m(x\,|\,\mu,\sigma,f) = \sigma^{-1} f\{\sigma^{-1}(x-\mu)\}, \quad x \in \Re, \quad \mu \in \Re, \quad \sigma > 0, \tag{2}$$

where $f(.)$ is some probability density in $\Re$, so that $f(y) \geq 0$, $\int_{\Re} f(y)\,dy = 1$. Interest lies in either the location parameter μ, the scale parameter σ, or some one-to-one function of these, and the likelihood function is

$$p(\boldsymbol{x}\,|\,\mu,\sigma) = \prod_{j=1}^{n} m(x_j\,|\,\mu,\sigma,f) = \sigma^{-n} \prod_{j=1}^{n} f\{\sigma^{-1}(x_j-\mu)\}. \tag{3}$$

Standard notation is used for the sample mean and the sample variance, respectively denoted by $\overline{x} = \sum_{j=1}^{n} x_j/n$ and $s^2 = \sum_{j=1}^{n}(x_j-\overline{x})^2/n$. Many conventional point estimators of the variance of location-scale models are members of the family of *affine invariant estimators*,

$$\tilde{\sigma}^2_{\nu} = \frac{ns^2}{\nu} = \frac{1}{\nu}\sum_{j=1}^{n}(x_j-\overline{x})^2, \quad \nu > 0\,. \tag{4}$$

In particular, with normal data, the MLE of the variance σ^2 is $s^2 = \tilde{\sigma}^2_n$, and the unbiased estimator is $\tilde{\sigma}^2_{n-1}$. More sophisticated estimators may sometimes be defined in terms of affine estimators; for instance, Stein (1964) and Brown (1968) estimators of the normal variance may respectively be written as

$$\tilde{\sigma}^2_{stein} = \min\left\{\tilde{\sigma}^2_{n+1},\, \tilde{\sigma}^2_{(n+2)/(1+z^2)}\right\}, \quad \tilde{\sigma}^2_{brown} = \min\left\{\tilde{\sigma}^2_{n-1},\, \tilde{\sigma}^2_{n/(1+z^2)}\right\},$$

where $z = \overline{x}/s$ is the standardized sample mean.

1.2 Contents

Section 2 provides a short review of intrinsic estimation, our approach to both point and region estimation. An information-theory based invariant loss function, the *intrinsic discrepancy*, is proposed as a reasonable general alternative to the conventional (non-invariant) quadratic loss. As is usually the case in modern literature, point estimation is described as a decision problem where the action space is the set of possible values for the quantity of interest; an intrinsic point estimator is then defined as the Bayes estimator which corresponds to the intrinsic loss and the appropriate reference prior. This provides a general *invariant* objective Bayes point estimator. Less conventionally, region estimation is also described as a decision problem where, for each p, the action space is the set of possible p-credible regions for the quantity of interest; a p-credible intrinsic region estimator is then defined as the lowest posterior loss p-credible region with respect to the intrinsic loss and the appropriate reference prior. This provides a

general *invariant* objective Bayes region estimator which always contains the intrinsic point estimator.

In Section 3 location-scale models are analyzed from this point of view. In particular, intrinsic point estimators and intrinsic region estimators are derived for the mean of a normal model, the variance of a normal model, and the scale parameter of a Cauchy model.

2 Intrinsic Estimation

2.1 The intrinsic discrepancy loss function

Point estimation of some parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is customarily described as a decision problem where the action space is the set $\mathcal{A} = \{\tilde{\boldsymbol{\theta}};\, \tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}\}$ of possible values of the vector of interest. Foundations dictate (see e.g., Bernardo and Smith, 1994, Ch. 2 and references therein) that to solve this decision problem it is necessary to specify a *loss function* $\ell\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\}$, such that $\ell\{\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}\} \geq 0$ and $\ell\{\boldsymbol{\theta}, \boldsymbol{\theta}\} = 0$, which describes, as a function of $\boldsymbol{\theta}$, the loss suffered from using $\tilde{\boldsymbol{\theta}}$ as a proxy for the unknown value of $\boldsymbol{\theta}$. The loss function is context specific, and should be chosen in terms of the anticipated uses of the estimate; however, a number of conventional loss functions have been suggested for those situations where no particular uses are envisaged, as in scientific inference. The simplest of these conventional loss functions (which typically ignore the presence nuisance parameters) is the ubiquitous *quadratic loss*, $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})$; the corresponding Bayes estimator, if this exists, is the *posterior mean*, $E[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$. Another common conventional loss function is the *zero-one loss*, defined as $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = 1$, if $\tilde{\boldsymbol{\theta}}$ does not belong to a ϵ-radius neighbourhood of $\boldsymbol{\theta}$, and zero otherwise; as $\epsilon \to 0$, the corresponding Bayes estimator converges to the posterior mode, $\mathrm{Mo}[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$. For details, see, e.g., Bernardo and Smith (1994, p. 257).

Example 1 *(Normal variance)* With the usual objective prior $\pi(\mu, \sigma) = \sigma^{-1}$, the (marginal) reference posterior density of σ is the square root inverted gamma

$$\pi(\sigma \,|\, \boldsymbol{x}) = \pi(\sigma \,|\, s, n) = \frac{n^{(n-1)/2}\, s^{n-1}}{2^{(n-3)/2}\, \Gamma[(n-1)/2]}\; \sigma^{-n}\, e^{-\frac{1}{2} n\, s^2/\sigma^2}, \quad n \geq 2. \tag{5}$$

The quadratic loss in terms of the variance, $\ell\{\tilde{\sigma}^2, \sigma^2\} = (\tilde{\sigma}^2 - \sigma^2)^2$, leads to $\mathrm{E}[\sigma^2 \,|\, \boldsymbol{x}] = \tilde{\sigma}^2_{n-3}$ (which obviously requires $n \geq 3$). Similarly, the quadratic loss in terms of the standard deviation, $\ell\{\tilde{\sigma}, \sigma\} = (\tilde{\sigma} - \sigma)^2$, yields

$$\mathrm{E}[\sigma \,|\, \boldsymbol{x}] = \sqrt{\frac{n}{2}}\, \frac{\Gamma[(n-2)/2]}{\Gamma[(n-1)/2]}\, s, \quad n \geq 3. \tag{6}$$

For moderate n values, the Stirling approximation of the Gamma functions in (6) produces $\mathrm{E}[\sigma \,|\, \boldsymbol{x}]^2 \approx \tilde{\sigma}^2_{n-5/2}$. Notice that, using the conventional quadratic loss function, the Bayes estimate of σ^2 is *not* the square of the Bayes estimate of σ. This lack of invariance is not an special feature of the quadratic loss; on the contrary, this is the case of most conventional loss functions. For instance, the use of the slightly more sophisticated standardized quadratic loss function on the variance,

$$\ell_{stq}(\tilde{\sigma}^2, \sigma^2) = [(\tilde{\sigma}^2/\sigma^2) - 1]^2 \tag{7}$$

yields (if $n \geq 2$, for $\pi(\sigma \,|\, \boldsymbol{x})$ to be proper)

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}^2/\sigma^2) - 1]^2 \, \pi(\sigma \,|\, \boldsymbol{x}) \, d\sigma = \frac{n\, s^2}{n+1} = \tilde{\sigma}^2_{n+1}, \tag{8}$$

which is also the minimum risk equivariant estimator (MRIE) of σ^2 under this loss, while the standardized quadratic loss function it terms of the standard deviation, $\ell(\tilde{\sigma}, \sigma) = [(\tilde{\sigma}/\sigma) - 1]^2$ yields (again, if $n \geq 2$)

$$\arg\min_{\tilde{\sigma}^2>0} \int_0^\infty [(\tilde{\sigma}/\sigma) - 1]^2 \, \pi(\sigma \,|\, \boldsymbol{x}) \, d\sigma = \sqrt{\frac{n}{2}} \, \frac{\Gamma[n/2]}{\Gamma[(n+1)/2]} \, s, \tag{9}$$

which is different from (6), and whose square is *not* (8). Similarly, for the zero-one loss in terms of σ^2, the Bayes estimator is the mode of the posterior distribution of σ^2, $\pi(\sigma^2 \,|\, \boldsymbol{x}) = \pi(\sigma \,|\, \boldsymbol{x})/(2\sigma)$, which is $\mathrm{Mo}(\sigma^2 \,|\, \boldsymbol{x}) = \tilde{\sigma}^2_{n+1}$, the same as (8), while the Bayes estimator for the zero-one loss in terms of σ is $\mathrm{Mo}(\sigma \,|\, \boldsymbol{x}) = s$, the MLE of σ, whose square is obviously not the same as (8). For further information on alternative point estimators of the normal variance, see Brewster and Zidek (1974) and Rukhin (1987).

As Example 1 dramatically illustrates, conventional loss functions are typically *not* invariant under reparametrization. As a consequence, the Bayes estimator $\boldsymbol{\phi}^*$ of a one-to-one transformation $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ of the original parameter $\boldsymbol{\theta}$ is not necessarily $\boldsymbol{\phi}(\boldsymbol{\theta}^*)$ and thus, for each loss function, one may produce as many *different* estimators of the same quantity as alternative parametrizations one is prepared to consider, a less than satisfactory situation. Indeed, scientific applications *require* this type of invariance. It would certainly be hard to argue that the best estimate of, say the age of the universe is θ^* but that the best estimate of the logarithm of that age is *not* $\log(\theta^*)$. Invariant loss functions are required to guarantee invariant estimators.

With no nuisance parameters, *intrinsic loss functions* (Robert, 1996), of the general form $\ell(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \ell\{p_x(. \,|\, \tilde{\boldsymbol{\theta}}), \, p_x(. \,|\, \boldsymbol{\theta})\}$ shift attention from the discrepancy between the estimate $\tilde{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}$, to the more relevant discrepancy between the statistical *models* they label, and they are always invariant under one-to-one reparametrization. The *intrinsic discrepancy*, introduced by Bernardo and Rueda (2002), is a particular intrinsic loss with specially attractive properties.

Definition 1 ***(Intrinsic Discrepancy)*** *The* intrinsic discrepancy *between two elements* $p_x(\cdot \,|\, \boldsymbol{\omega}_1)$ *and* $p_x(\cdot \,|\, \boldsymbol{\omega}_2)$ *of the parametric family of distributions* $\mathcal{F} = \{p_x(\cdot \,|\, \boldsymbol{\omega}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}\}$*, is*

$$\begin{aligned} \delta_x(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2) &= \delta\{p_x(. \,|\, \boldsymbol{\omega}_1),\ p_x(. \,|\, \boldsymbol{\omega}_2)\} = \min\{\kappa_x(\boldsymbol{\omega}_1 \,|\, \boldsymbol{\omega}_2),\ \kappa_x(\boldsymbol{\omega}_2 \,|\, \boldsymbol{\omega}_1)\}, \\ \kappa_x(\boldsymbol{\omega}_j \,|\, \boldsymbol{\omega}_i) &= \int_{\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_i)} p_x(\boldsymbol{x} \,|\, \boldsymbol{\omega}_i) \log \frac{p_x(\boldsymbol{x} \,|\, \boldsymbol{\omega}_i)}{p_x(\boldsymbol{x} \,|\, \boldsymbol{\omega}_j)}\, d\boldsymbol{x}, \end{aligned}$$

The intrinsic discrepancy $\delta_x\{\mathcal{F}_1, \mathcal{F}_2\}$ *between two subsets* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *of* $\mathcal{F}$ *is the minimum intrinsic discrepancy between its elements,*

$$\delta_x(\mathcal{F}_1, \mathcal{F}_2) = \min_{\boldsymbol{\omega}_1 \in \mathcal{F}_1,\, \boldsymbol{\omega}_2 \in \mathcal{F}_2} \delta\{p_x(. \,|\, \boldsymbol{\omega}_1),\ p_x(. \,|\, \boldsymbol{\omega}_2)\}$$

Thus, the intrinsic discrepancy $\delta(\boldsymbol{\omega}_1, \boldsymbol{\omega}_2)$ between two parameter values $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ is the minimum Kullback-Leibler directed logarithmic divergence (Kullback and Leibler, 1951) between the distributions $p_x(. \,|\, \boldsymbol{\omega}_1)$ and $p_x(. \,|\, \boldsymbol{\omega}_2)$ which they label. Notice that this is obviously independent of the particular parametrization chosen to describe the distributions. The intrinsic discrepancy is a divergence measure in the class $\mathcal{F}$; indeed, (i) it is symmetric, (ii) it is non-negative and (iii) it is zero if, and only if, $p_x(\boldsymbol{x} \,|\, \boldsymbol{\omega}_1) = p_x(\boldsymbol{x} \,|\, \boldsymbol{\omega}_2)$ almost everywhere. Notice that in Definition 1 the possible dependence of the sampling space $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}(\boldsymbol{\omega})$ on the parameter value $\boldsymbol{\omega}$ is explicitly allowed, so that the intrinsic discrepancy may be used with non-regular models where the support $\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_1)$ of, say, $p_x(. \,|\, \boldsymbol{\omega}_1)$ may be strictly smaller than the support $\boldsymbol{\mathcal{X}}(\boldsymbol{\omega}_2)$ of $p_x(. \,|\, \boldsymbol{\omega}_2)$.

The intrinsic discrepancy is also invariant under one-to-one transformations of the random vector $\boldsymbol{x}$. Moreover, directed logarithmic divergences are *additive* with respect to conditionally independent observations. Consequently, if $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is a random sample from, say $q_x(\cdot \,|\, \boldsymbol{\omega})$ so that the probability model is $p(\boldsymbol{x} \,|\, \boldsymbol{\omega}) = \prod_{i=1}^{n} q(x_j \,|\, \boldsymbol{\omega})$, then the intrinsic discrepancy $\delta_x\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2\}$ between $p_x(\cdot \,|\, \boldsymbol{\omega}_1)$ and $p_x(\cdot \,|\, \boldsymbol{\omega}_2)$ is simply $n\, \delta_x\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2\}$, that is, n times the intrinsic discrepancy between $q_x(\cdot \,|\, \boldsymbol{\omega}_1)$ and $q_x(\cdot \,|\, \boldsymbol{\omega}_2)$.

In the context of point estimation, the intrinsic discrepancy leads naturally to the (invariant) intrinsic discrepancy loss $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ defined as the intrinsic discrepancy between the assumed model $p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ and its closest approximation within the set $\{p_x(\cdot \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}\}$ of all models with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Definition 2 ***(Intrinsic discrepancy loss)*** *Consider the family of probability distributions* $\mathcal{F} = \{p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}, \boldsymbol{\lambda})\}$*. The* intrinsic discrepancy loss *from using* $\tilde{\boldsymbol{\theta}}$ *as a proxy for* $\boldsymbol{\theta}$ *is*

$$\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \delta_x\{(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}), (\boldsymbol{\theta}, \boldsymbol{\lambda})\},$$

the intrinsic discrepancy between $p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ *and the set* $\{p_x(\cdot \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}\}$*.*

Notice that the value of $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ does *not* depend on the particular parametrization chosen to describe the problem. Indeed, for any one-to-one reparametrizations $\boldsymbol{\phi} = \boldsymbol{\phi}(\boldsymbol{\theta})$ and $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\lambda})$,

$$\delta_x\{\tilde{\boldsymbol{\phi}}, (\boldsymbol{\phi}, \boldsymbol{\psi})\} = \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} \tag{10}$$

so that, as one should surely require, the loss suffered from using $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}(\tilde{\boldsymbol{\theta}})$ as a proxy for $\boldsymbol{\phi}(\boldsymbol{\theta})$ is precisely the same as the loss suffered from using $\tilde{\boldsymbol{\theta}}$ as a proxy for $\boldsymbol{\theta}$, and this is true for any parametrization of the nuisance parameter vector.

Under frequently met regularity conditions, the two minimizations required in Definition 2 may be interchanged. This makes analytical derivation of the intrinsic loss considerably simpler.

Theorem 1 ***(Computation of the intrinsic discrepancy loss)*** *Let $\mathcal{F}$ be a parametric family of probability distributions*

$$\mathcal{F} = \{p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})\},$$

with convex support $\boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ for all $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Then,

$$\begin{aligned} \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} &= \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \min\left\{\kappa_x\{\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \kappa_x\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}\}\right\} \\ &= \min\left\{\inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \kappa_x\{\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}\},\ \inf_{\tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda}} \kappa_x\{\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}\}\right\} \end{aligned}$$

Proof. This follows from the fact that, if $\boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a convex set, then the two directed logarithmic divergences involved in the definition are convex functionals. For details, see Juárez (2004). □

Example 2 *(Normal variance, continued)* Consider $\boldsymbol{x} = \{x_1, \ldots, x_n\}$. The directed logarithmic divergence of $\prod_{i=1}^n \mathrm{N}(x_j \,|\, \tilde{\mu}, \tilde{\sigma})$ from $\prod_{i=1}^n \mathrm{N}(x_j \,|\, \mu, \sigma)$ is

$$\begin{aligned} \kappa_x\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\} &= n \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu, \sigma) \log\left[\frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \tilde{\mu}, \tilde{\sigma})}\right] dx \\ &= \frac{n}{2}\left[\frac{\sigma^2}{\tilde{\sigma}^2} - 1 - \log\frac{\sigma^2}{\tilde{\sigma}^2} + \frac{(\mu - \tilde{\mu})^2}{\tilde{\sigma}^2}\right]. \end{aligned} \tag{11}$$

This is minimized when $\tilde{\mu} = \mu$, to yield

$$\inf_{\tilde{\mu} \in \mathbb{R}} \kappa_x\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\} = \frac{n}{2}\, g\left(\frac{\sigma^2}{\tilde{\sigma}^2}\right) = \frac{n}{2}\, g\,(\phi)\,,$$

where $\phi = \tilde{\sigma}^2/\sigma^2$, and $g(.)$ is the *linlog* function defined by

$$g(t) = (t-1) - \log t, \quad t > 0. \tag{12}$$

Notice that $g(t) \geq 0$ and $g(t) = 0$ if (and only if) $t = 1$. This follows from the fact that $g(t)$ is the absolute distance between $\log t$ and its tangent at $t = 1$.

Exchanging the roles of $(\tilde{\mu}, \tilde{\sigma})$ and (μ, σ), it is similarly found that $\kappa_x\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\}$ is also minimized when $\tilde{\mu} = \mu$, to yield

$$\inf_{\tilde{\mu} \in \mathbb{R}} \kappa_x\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\} = \frac{n}{2}\, g\left(\frac{\tilde{\sigma}^2}{\sigma^2}\right) = \frac{n}{2}\, g\left(\frac{1}{\phi}\right) \cdot$$

Moreover, $g(t) < g(1/t)$ if, and only if, $t < 1$ and hence, using Theorem 1,

$$\delta_x\{\tilde{\sigma}, (\mu, \sigma)\} \;=\; \delta_x\{\phi\} = \begin{cases} \frac{n}{2}\, g(\phi) & \text{if} \quad \phi < 1, \\ \frac{n}{2}\, g(1/\phi) & \text{if} \quad \phi \geq 1, \end{cases} \qquad \phi = \frac{\tilde{\sigma}^2}{\sigma^2} \cdot \tag{13}$$

Thus, for fixed n, the intrinsic discrepancy loss $\delta_x\{\tilde{\sigma}, (\mu, \sigma)\}$ only depends on the ratio $\phi = \tilde{\sigma}^2/\sigma^2$. The intrinsic discrepancy loss is closely related to the (also invariant) *entropy* loss,

$$\ell_{ent}\{\tilde{\sigma}, \sigma\} = \ell_{ent}\{\phi\} = \int_{-\infty}^{\infty} \mathrm{N}(x \,|\, \mu, \sigma) \log\left[\frac{\mathrm{N}(x \,|\, \mu, \sigma)}{\mathrm{N}(x \,|\, \mu, \tilde{\sigma})}\right] dx = \frac{1}{2}\, g(\phi), \tag{14}$$

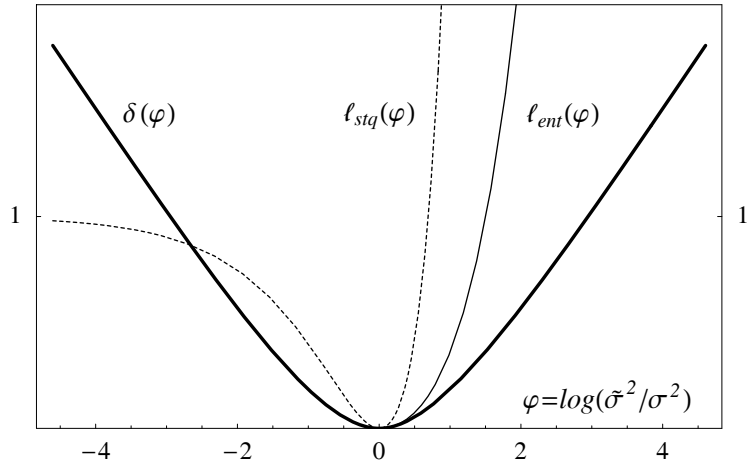

Figure 1: *Intrinsic discrepancy loss (solid line), entropy loss (continuous line), and standardized quadratic loss (dotted line) for point estimation of the normal variance, as a function of* $\psi = \log(\tilde{\sigma}^2/\sigma^2)$.

which Brown (1990) attributes to Stein. Except for the proportionality constant n (which does not affect estimation), the entropy loss (14) is the same as the intrinsic discrepancy loss (13) whenever $\tilde{\sigma} < \sigma$. Indeed, the intrinsic discrepancy loss may be seen as a symmetrized version of the entropy loss.

Notice that, for all values of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$, $\delta_x\{\phi\} = \delta_x\{1/\phi\}$; hence, the intrinsic loss *equally* penalizes overestimation and underestimation. In sharp contrast, both the entropy loss and the often recommended standardized quadratic loss function, which is also a function of the ratio ϕ,

$$\ell_{stq}(\tilde{\sigma}^2, \sigma^2) = [(\tilde{\sigma}^2/\sigma^2) - 1]^2 = (\phi - 1)^2,$$

clearly underpenalize small estimators, thus yielding estimators of the variance which are too small. This is illustrated in Figure 1, where the functions $\delta_x(\phi)$ (for $n = 1$), $\ell_{ent}\{\phi\}$, and $\ell_{stq}(\phi)$ are all represented as a function of $\varphi = \log \phi$. More conventional loss functions, as the usual quadratic loss,

$$\ell_{quad}(\tilde{\sigma}^2, \sigma^2) = [\tilde{\sigma}^2 - \sigma^2]^2 = \sigma^4(\phi - 1)^2,$$

are not even invariant with respect to affine transformations. All this led Stein (1964, p. 156) to write "I find it hard to take the problem of estimating σ^2 with quadratic loss very seriously".

2.2 Reference posterior expectation

Given data $\boldsymbol{x}$ generated by $p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, a situation with no prior information about the value of $\boldsymbol{\theta}$ is formally described by the *reference prior* $\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ which corresponds to the model $p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$ when $\boldsymbol{\theta}$ is the quantity of interest (Bernardo, 1979; Berger and Bernardo, 1992; Bernardo, 2005a). In this case, all available information about $\boldsymbol{\theta}$ is encapsulated it its (marginal) reference posterior distribution, $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \int_{\Lambda} \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\lambda}$ where, by Bayes theorem, $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. If numerical summaries of the information encapsulated in $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})$ are further required in the form of either point or region estimators of $\boldsymbol{\theta}$ under some specified loss function $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$, then the reference posterior expected loss

$$l(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \int_{\Theta} \int_{\Omega} \ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}$$

(from using $\tilde{\boldsymbol{\theta}}$ as a proxy for $\boldsymbol{\theta}$) has to be evaluated. In view of the arguments given above, attention will focus on the intrinsic discrepancy reference expected loss, or *intrinsic expected loss*, for short.

Definition 3 ***(Intrinsic expected loss)*** *Consider the parametric family of probability distributions*

$$\mathcal{F} = \{p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\theta}, \boldsymbol{\lambda}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta},\ \boldsymbol{\lambda} \in \boldsymbol{\Lambda},)\}. \tag{15}$$

The intrinsic expected loss *from using $\tilde{\boldsymbol{\theta}}$ given data $\boldsymbol{x}$, denoted $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$, is the posterior expectation of the intrinsic discrepancy loss, $\delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}$ with respect to the joint reference posterior, $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})$,*

$$d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \int_{\boldsymbol{\Theta}} \int_{\boldsymbol{\Omega}} \delta_x\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\}\, \pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{\lambda}, \tag{16}$$

where $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda} \,|\, \boldsymbol{x}) \propto p(\boldsymbol{x} \,|\, \boldsymbol{\theta}, \boldsymbol{\omega})\, \pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$, and $\pi(\boldsymbol{\lambda} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$ is the (joint) reference prior when $\boldsymbol{\theta}$ is the quantity of interest.

The function $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ measures the posterior expected loss from using $\tilde{\boldsymbol{\theta}}$ as a proxy for the unknown value of $\boldsymbol{\theta}$, in terms of the expected intrinsic discrepancy between the assumed model, $p_x(\cdot \,|\, \boldsymbol{\theta}, \boldsymbol{\lambda})$, and the class

$$\mathcal{F}_{\tilde{\boldsymbol{\theta}}} = \{p(\boldsymbol{x} \,|\, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\lambda}}),\ \tilde{\boldsymbol{\lambda}} \in \boldsymbol{\Lambda},\ \boldsymbol{x} \in \boldsymbol{\mathcal{X}}(\boldsymbol{\omega}, \boldsymbol{\lambda}) \tag{17}$$

of those models in $\mathcal{F}$ for which $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

The intrinsic expected loss provides an objective measure of the *compatibility* of the value $\tilde{\boldsymbol{\theta}}$ with the observed data $\boldsymbol{x}$, with a nice interpretation in terms of likelihood ratios. Indeed, it immediately follows from Definitions 1, 2 and 3, that $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ is the posterior expectation of the minimum expected log-likelihood ratio between the true model and the closest model for which $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$. For instance, if $d(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) = \log 100$, then data $\boldsymbol{x}$ are expected to be about 100 times more likely under the true (unknown) model than under any model within this family with $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

Example 3 *(Normal variance, continued)* In Example 2 the intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy for σ was seen to be a function $\delta_x(\phi)$ of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$ (Equation 13). Changing variables in (5), the reference posterior of ϕ is

$$\pi(\phi \,|\, \boldsymbol{x}) = \mathrm{Ga}(\phi \,\Big|\, \frac{n-1}{2}\,,\, \frac{n\, s^2}{2\tilde{\sigma}^2}) \cdot \tag{18}$$

Hence, the intrinsic expected loss from using $\tilde{\sigma}$ as a proxy for σ is

$$d(\tilde{\sigma} \,|\, \boldsymbol{x}) = d(\tilde{\sigma} \,|\, s^2, n) = \int_0^\infty \delta_x(\phi)\, \mathrm{Ga}(\phi \,|\, \frac{n-1}{2}\,,\, \frac{n\, s^2}{2\tilde{\sigma}^2})\, d\phi \tag{19}$$

which may easily be computed by one-dimensional numerical integration. Good analytical approximations will however be provided is Section 3. As a numerical

illustration, a random sample of size $n = 12$ was simulated from a normal distribution with $\mu = 5$ and $\sigma = 2$, yielding $\overline{x} = 4.214$ and $s = 2.071$. The corresponding intrinsic expected loss $d(\tilde{\sigma} \,|\, \boldsymbol{x})$, represented in the lower panel of Figure 2, is locally convex around a unique minimum.

2.3 Intrinsic Point and Region Estimation

Bayes estimates are, by definition, those which minimize the expected posterior loss. The *intrinsic estimate* is the Bayes estimate which corresponds to the intrinsic discrepancy loss and the reference posterior distribution, i.e., that value $\tilde{\boldsymbol{\theta}}_{int}(\boldsymbol{x}) \in \boldsymbol{\Theta}$ which minimizes the intrinsic expected loss. Formally,

Definition 4 ***(Intrinsic point estimator)*** *Consider again the parametric family of probability distributions $\mathcal{F}$ defined by* (15). *An* intrinsic estimator *of $\boldsymbol{\theta}$ is a value*

$$\tilde{\boldsymbol{\theta}}_{int}(\boldsymbol{x}) = \min_{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta}} d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\},$$

which minimizes the intrinsic discrepancy reference posterior loss (16).

Under general regularity conditions, the intrinsic expected loss $d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$ is locally convex near its absolute minimum and, therefore, the intrinsic estimate is typically exists and it is unique. Moreover, since both the intrinsic loss function and the reference prior are invariant under one-to-one reparametrization, the intrinsic estimator $\tilde{\boldsymbol{\psi}}_{int}(\boldsymbol{x})$ of any one-to-one function $\boldsymbol{\psi}(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$ will simply be $\tilde{\boldsymbol{\psi}}_{int} = \boldsymbol{\psi}(\tilde{\boldsymbol{\theta}}_{int})$. For more details on intrinsic estimation, see Bernardo and Juárez (2003).

Bayesian region estimation is typically based on posterior credible regions, i.e., sets of $\boldsymbol{\theta}$ values with pre-specified posterior probabilities. However, for any fixed p, there are typically infinitely many p-credible regions. In most cases, these are chosen to be either highest posterior density (HPD) regions, or probability centred regions.

It is well known that the ubiquitous *highest posterior density* (HPD) regions are *not* consistent under reparametrization. Thus, if $\boldsymbol{\phi}\{\boldsymbol{\theta}\}$ is a one-to-one function of $\boldsymbol{\theta}$, the image of a HPD p-credible region of $\boldsymbol{\theta}$ will *not* generally be HPD for ϕ. Thus, if C_p is a HPD p-credible set estimate for, say, the perihelion of the Earth, $\log(C_p)$ will *not* be a HPD p-credible set estimate for its logarithm. This suggests that highest posterior density may not be a good criterion for set estimation in scientific inference.

In *one-dimensional* problems, one may define *probability centred* credible intervals, and these are invariant under reparametrization. Indeed, the probability centred p-credible interval of a real-valued quantity of interest θ is defined by the $(1 - p)/2$ and $(1 + p)/2$ quantiles of its posterior distribution $\pi(\theta \,|\, \boldsymbol{x})$, and this is invariant under one-to-one reparametrizations, since all quantiles are invariant. However, the notion cannot be uniquely extended to multidimensional problems and, even in one-dimensional

problems, their use may less than satisfactory as, for instance, in those situations where the posterior density is monotonously decreasing within its support.

Whenever a loss structure has been established, foundations dictate that values with smaller expected loss are to be preferred. Thus, for any loss function $\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\omega})\}$ is it natural to define p-credible *lowest posterior loss* (LDL) region estimators (Bernardo, 2005b) as those p-credible regions which contain $\tilde{\boldsymbol{\theta}}$ values whose expected loss $l(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x})$ (Eq. 15), is smaller than that of any $\tilde{\boldsymbol{\theta}}$ values outside the region.

In particular, if the loss function is quadratic, so that

$$\ell\{\tilde{\boldsymbol{\theta}}, (\boldsymbol{\theta}, \boldsymbol{\lambda})\} = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}),$$

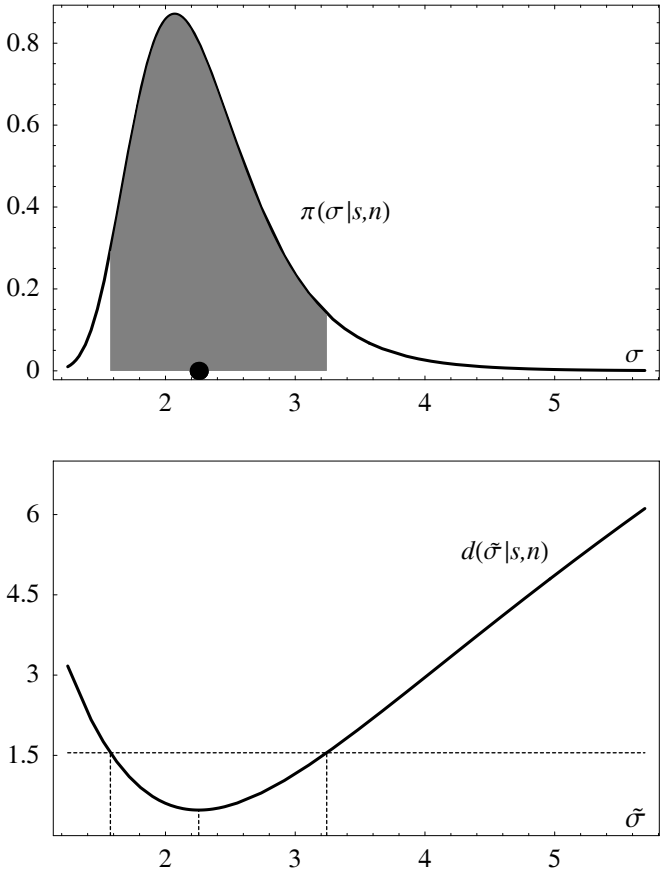

Figure 2: *Reference posterior density of the standard deviation σ of a normal distribution (upper panel), and intrinsic expected loss from using $\tilde{\sigma}$ as a proxy for σ (lower panel), given a random sample $\boldsymbol{x}$ of size $n = 12$ with standard deviation $s = 2.071$. The intrinsic estimate (solid dot) is $\tilde{\sigma}_{int} = 2.256$; the 0.90 intrinsic credible region (shaded region) is $C^{int}_{0.90} = (1.575, 3.243)$.*

the expected loss is

$$\begin{aligned} l(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}) &= \int_{\boldsymbol{\Theta}} (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^t (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} \\ &= (\tilde{\boldsymbol{\theta}} - \mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{x}])^t (\tilde{\boldsymbol{\theta}} - \mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{x}]) + \mathrm{Var}[\boldsymbol{\theta} \,|\, \boldsymbol{x}]. \end{aligned}$$

Hence, with quadratic loss, the lowest posterior loss p-credible region consists of those $\tilde{\boldsymbol{\theta}}$ values with the smallest Euclidean distance to the posterior mean $\mathrm{E}[\boldsymbol{\theta} \,|\, \boldsymbol{x}]$. Notice that these LDL p-credible regions are *not* invariant under reparametrization.

To obtain LDL invariant region estimators the loss function used must be invariant under one-to-one reparametrization. The arguments mentioned above suggest the use of the intrinsic discrepancy loss. The p-credible *intrinsic region estimator* is the lowest posterior loss p-credible region which corresponds to the intrinsic discrepancy loss.

Definition 5 ***(Intrinsic region estimator)*** *Consider once more the parametric family of probability distributions $\mathcal{F}$ defined by* (15). *An* intrinsic p-credible region *for $\boldsymbol{\theta}$ is a subset $C_p^{int} = C_p^{int}(\boldsymbol{x}, \boldsymbol{\Theta})$ of $\boldsymbol{\Theta}$ such that*

$$\int_{C_p^{int}} \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x})\, d\boldsymbol{\theta} = p, \quad \forall \tilde{\boldsymbol{\theta}}_i \in C_p^{int},\ \forall \tilde{\boldsymbol{\theta}}_j \notin C_p^{int},\ d\{\tilde{\boldsymbol{\theta}}_i \,|\, \boldsymbol{x}\} < d\{\tilde{\boldsymbol{\theta}}_j \,|\, \boldsymbol{x}\},$$

where, again, $d\{\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{x}\}$ is the intrinsic expected loss (16).

Intrinsic credible regions are typically unique and, since they are based in the invariant intrinsic discrepancy loss, they are consistent under one-to-one reparametrization. Thus, if $\boldsymbol{\psi}(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, the image $C_p^{int}(\boldsymbol{x}, \boldsymbol{\Psi}) = \boldsymbol{\psi}\{C_p^{int}(\boldsymbol{x}, \boldsymbol{\Theta})\}$ of an intrinsic p-credible region for $\boldsymbol{\theta}$ is an intrinsic p-credible region for $\boldsymbol{\phi}$. For more details on intrinsic region estimation, see Bernardo (2005b).

Example 4 *(Normal variance, continued)* Numerical minimization of the intrinsic expected loss (19) in Example 3 immediately yields the intrinsic estimator of of the standard deviation σ. This is

$$\sigma^*(\boldsymbol{x}) = \sigma^*(n, s) = \arg\min_{\tilde{\sigma}>0}\ d(\tilde{\sigma} \,|\, \boldsymbol{x}) = 2.256, \tag{20}$$

and it is marked with a solid dot in the top panel of Figure 2. Since intrinsic estimation is invariant, the intrinsic estimates of σ^2 or $\log \sigma$ are respectively $(\sigma^*)^2$ and $\log(\sigma^*)$.

Moreover, the intrinsic p-credible interval for σ is given by $C_p^{int} = (\sigma_0, \sigma_1)$, where $\{\sigma_0, \sigma_1\}$ is the unique solution to the equations system

$$\begin{cases} d(\sigma_0 \,|\, \boldsymbol{x}) = d(\sigma_1 \,|\, \boldsymbol{x}) \\ \int_{\sigma_0}^{\sigma_1} \pi(\sigma \,|\, \boldsymbol{x})\, d\sigma = p \end{cases}$$

For instance, with $p = 0.90$ this yields $C^{int}_{0.90} = (1.575, 3.243)$, the shaded region in the top panel of Figure 2. Since intrinsic region estimation is also invariant under reparametrization, the intrinsic p-credible intervals for σ^2 or $\log\sigma$ will respectively be $(C^{int}_p)^2$ and $\log(C^{int}_p)$.

3 Intrinsic estimation in location-scale models

This section analyses intrinsic point and region estimation of the parameters μ and σ (or arbitrary one-to-one functions of these) of variation-independent location-scale models.

3.1 Reference analysis of location-scale models

The likelihood function $p(\boldsymbol{x}\,|\,\mu,\sigma,f)$ which corresponds to a random sample $\boldsymbol{x} = \{x_1,\ldots,x_n\}$ from a location-scale model $m(x\,|\,\mu,\sigma,f)$ of the form (2), is given by Equation (3). This will typically have a unique maximum, denoted $(\hat{\mu},\hat{\sigma})$, which not always has a simple analytical expression.

Under appropriate regularity conditions (see, e.g., Bernardo and Smith, 1994, Sec. 5.3 and references therein) the joint posterior distribution of μ and σ will be asymptotically normal with mean $(\hat{\mu},\hat{\sigma})$ and covariance matrix

$$V(\hat{\mu},\hat{\sigma},n) = n^{-1}\,F^{-1}(\hat{\mu},\hat{\sigma}) = (\tilde{\sigma}^2/n)\,A^{-1}(f) \tag{21}$$

where $F(\mu,\sigma)$ is Fisher's information matrix which, in location-scale models, is always of the form $F(\mu,\sigma) = \sigma^{-2}A(f)$, where $A(f)$ is a 2×2 matrix which depends on the probability density $f(\cdot)$, but not on μ or σ. As a consequence, if the parameter of interest is either μ (or a one-to-one function of μ) or σ (or a one-to-one function of σ) with independent variation, then (Fernández and Steel, 1999, Th. 1), under regularity conditions sufficient to guarantee posterior asymptotic normality, and variation independence of μ and σ, the joint reference prior is independent of the function $f(\cdot)$ and, in terms of μ and σ, is given by

$$\pi(\mu)\,\pi(\sigma\,|\,\mu) = \pi(\sigma)\,\pi(\mu\,|\,\sigma) = \sigma^{-1}. \tag{22}$$

Using Bayes theorem, the corresponding joint reference posterior is

$$\pi(\mu,\sigma\,|\,\boldsymbol{x}) = \sigma^{-(n+1)}\prod_{j=1}^{n} f\{\sigma^{-1}(x_j-\mu)\}, \tag{23}$$

which is typically proper for all $n \geq 2$. In particular, it is proper for all $n \geq 2$ whenever

$f(.)$ is either a standard normal or a scale mixture of standard normals, what includes Student models.

In the normal case, with $f(x) = \mathrm{N}(x \,|\, 0, 1)$, the joint posterior (23) becomes

$$\pi(\mu, \sigma \,|\, \overline{x}, s, n) = \mathrm{N}(\mu \,|\, \overline{x}, \sigma/\sqrt{n})\, \pi(\sigma \,|\, s, n),$$

where $\pi(\sigma \,|\, s, n)$ is the square root inverted gamma given by (5). The corresponding marginal reference posterior of the precision $\lambda = \sigma^{-2}$ is found to be $\pi(\lambda \,|\, \boldsymbol{x}) = \mathrm{Ga}(\lambda \,|\, (n-1)/2, (ns^2)/2)$ and, thus,

$$\mathrm{E}[\lambda \,|\, \boldsymbol{x}] = \frac{n-1}{n\, s^2}\,, \quad \mathrm{Var}[\lambda \,|\, \boldsymbol{x}] = \frac{2(n-1)}{n^2\, s^4}\,. \tag{24}$$

The marginal reference posterior of μ is the Student distribution

$$\pi(\mu \,|\, \boldsymbol{x}) = \mathrm{St}\Big(\mu \,\Big|\, \overline{x},\ \frac{s}{\sqrt{n-1}}\,, n\Big) \propto \Big(1 + \frac{(\mu - \overline{x})^2}{s^2}\Big)^{-n/2}. \tag{25}$$

For details see, for instance, Bernardo and Smith (1994, Sec. 5.4).

3.2 Intrinsic estimation of the normal mean

As stated in Example 2 (Eq. 11), the directed divergence $\kappa\{\mu_j, \sigma_j \,|\, \mu_i, \sigma_i\}$, of $\mathrm{N}(x \,|\, \mu_j, \sigma_j)$ from $\mathrm{N}(x \,|\, \mu_i, \sigma_i)$, is

$$\kappa\{\mu_j, \sigma_j \,|\, \mu_i, \sigma_i\} = \frac{1}{2}\Big\{\frac{\sigma_i^2}{\sigma_j^2} - 1 - \log\frac{\sigma_i^2}{\sigma_j^2} + \frac{(\mu_i - \mu_j)^2}{\sigma_j^2}\Big\}.$$

As a function of $\tilde{\sigma}$, the directed divergence $\kappa\{\tilde{\mu}, \tilde{\sigma} \,|\, \mu, \sigma\}$ is minimized when $\tilde{\sigma}^2$ takes the value $\tilde{\sigma}^2_{min} = (\mu - \mu_i)^2 + \sigma^2$, and substitution yields

$$\kappa\{\tilde{\mu}, \tilde{\sigma}_{min} \,|\, \mu, \sigma\} = \frac{1}{2}\log\left[1 + \frac{(\mu - \tilde{\mu})^2}{\sigma^2}\right].$$

Similarly, the directed divergence $\kappa\{\mu, \sigma \,|\, \tilde{\mu}, \tilde{\sigma}\}$ is minimized, as a function of $\tilde{\sigma}$, when $\tilde{\sigma} = \sigma$, and substitution now yields

$$\kappa\{\mu, \sigma \,|\, \tilde{\mu}, \sigma\} = \frac{1}{2}\,\frac{(\mu - \tilde{\mu})^2}{\sigma^2}\,.$$

Hence, making use of Theorem 1 and the fact that, for all $x > 0$, $\log(1 + x) \leq x$, the intrinsic discrepancy loss $\delta\{\tilde{\mu}, (\mu, \sigma)\}$ from using $\tilde{\mu}$ as a proxy for μ with a normal sample of size n is

$$\delta\{\tilde{\mu}, (\mu, \sigma)\} = \delta\{\theta^2\} = \frac{n}{2} \log[1 + \frac{\theta^2}{n}], \quad \theta = \theta(\tilde{\mu}, \mu, \sigma) = \frac{\mu - \tilde{\mu}}{\sigma/\sqrt{n}}, \tag{26}$$

which only depends on the number θ of standard deviations which separate $\tilde{\mu}$ from μ. Figure 3 represents the intrinsic loss function (26), as a function of θ, for several values of n.

As one might expect, $\delta\{\tilde{\mu}, (\mu, \sigma)\}$ increases with $|\theta|$. The dependence is essentially quadratic in a neighbourhood of zero, but shows a very reasonable concavity in regions where $|\theta|$ is large.

Using Definition 3, the intrinsic discrepancy reference expected loss $d(\tilde{\mu} \,|\, \boldsymbol{x})$ may be written in terms of the reference posterior of θ; indeed,

$$\begin{aligned} d(\tilde{\mu} \,|\, \boldsymbol{x}) &= \int_0^\infty \int_{-\infty}^\infty \delta\{\tilde{\mu}, (\mu, \sigma)\}\, \pi(\mu, \sigma \,|\, \boldsymbol{x})\, d\mu\, d\sigma \\ &= \int_0^\infty \frac{n}{2} \log\left[1 + \frac{\theta^2}{n}\right] \pi(\theta \,|\, \boldsymbol{x})\, d\theta. \end{aligned} \tag{27}$$

But $\theta = (\tilde{\mu} - \mu)/(\sigma/\sqrt{n})$ may be written as $a + \beta$ where, as a function of μ and σ, $\beta = (\mu - \overline{x})/(\sigma/\sqrt{n})$ has a standard normal reference posterior, and a is the constant $a = (\overline{x} - \tilde{\mu})/(\sigma/\sqrt{n})$. Hence, the conditional posterior distribution of θ^2 given σ is noncentral χ^2 with one degree of freedom and non centrality parameter a^2,

$$\pi(\theta^2 \,|\, \boldsymbol{x}, \sigma) = \chi^2(\theta^2 \,|\, 1, a^2), \quad a^2 = n\, \frac{(\overline{x} - \tilde{\mu})^2}{\sigma^2}.$$

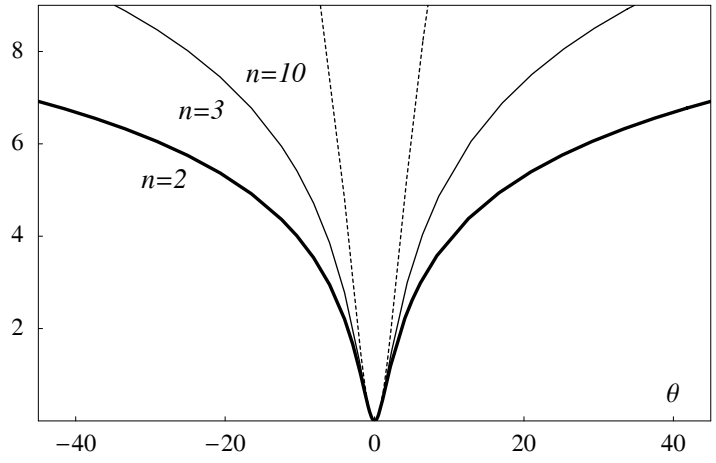

Figure 3: *Intrinsic discrepancy loss for estimation of the normal mean as a function of the number* $\theta = (\tilde{\mu} - \mu)/(\sigma/\sqrt{n})$ *of standard deviations which separates* $\tilde{\mu}$ *from* μ*, for* $n = 2$*,* $n = 3$*, and* $n = 10$*.*

It follows that the intrinsic expected loss $d(\tilde{\mu} \,|\, \boldsymbol{x})$ only depends on $\tilde{\mu}$ through $(\overline{x} - \tilde{\mu})^2$, and increases with $(\overline{x} - \tilde{\mu})^2$; therefore, the intrinsic estimator of μ is

$$\tilde{\mu}_{int}(\boldsymbol{x}) = \arg\min_{\tilde{\mu} \in \Re} \; d(\tilde{\mu} \,|\, \boldsymbol{x}) = \arg\min_{\tilde{\mu} \in \Re} \; (\overline{x} - \tilde{\mu})^2 = \overline{x}.$$

Moreover, $d(\tilde{\mu} \,|\, \boldsymbol{x})$ is symmetric around $\overline{x}$ and, hence, all intrinsic credible regions must be centered at $\overline{x}$. In view of (25), this implies that the intrinsic the p-credible regions are just the usual Student-t HPD p-credible intervals

$$C_p^{int}(\boldsymbol{x}) = \left\{ \tilde{\mu};\; \overline{x} - q_{p,n}\, s/\sqrt{n-1} \le \tilde{\mu} \le \overline{x} + q_{p,n}\, s/\sqrt{n-1} \right\}, \tag{28}$$

where $q_{p,n}$ is the $(p+1)/2$ quantile of a standard Student-t with $n-1$ degrees of freedom.

It immediately follows from (28) that C_p^{int} consist of the set of $\tilde{\mu}$ values such that $(\overline{x} - \tilde{\mu})/(s/\sqrt{n-1})$ belongs to a probability p centred interval of a standard Student-t with $n-1$ degrees of freedom. But, as a function of the data $\boldsymbol{x}$, the sampling distribution of

$$t(\boldsymbol{x}) = (\overline{x} - \mu)/(s/\sqrt{n-1}) \tag{29}$$

is also a standard Student-t with $n-1$ degrees of freedom. Hence, for all sample sizes, the *expected coverage under sampling* of the p-credible intervals (28) is *exactly* p, and the intrinsic credible regions are exact frequentist confidence intervals.

A simple asymptotic approximation to $d(\tilde{\mu} \,|\, \boldsymbol{x})$, which provides a direct measure in a log-likelihood ratio scale of the expected loss associated to the use of $\tilde{\mu}$, may easily be obtained. Indeed, a variation of the delta method shows that, under appropriate regularity conditions, the expectation of some function $y = g(x)$ of a random quantity x with mean μ_x and variance σ_x^2 may be approximated by

$$E[g(x)] \approx g\left[\mu_x + \frac{\sigma_x^2}{2} \frac{g''(\mu_x)}{g'(\mu_x)}\right]. \tag{30}$$

On the other hand, the conditional posterior mean of θ^2 is $1 + a^2$, and its conditional posterior variance is $2 + 4a^2$; but $\mathrm{E}[\sigma^{-2} \,|\, \boldsymbol{x}] = \mathrm{E}[\lambda \,|\, \boldsymbol{x}] = (n-1)/(ns^2)$ (Eq. 24) and hence, the unconditional posterior mean and variance of $\theta^2(\tilde{\mu})$ are, respectively,

$$\mathrm{E}[\theta^2 \,|\, \boldsymbol{x}] = 1 + t^2, \quad \mathrm{Var}[\theta^2 \,|\, \boldsymbol{x}] = 2 + 4\,t^2,$$

both functions of the conventional t statistic (29). Using these in (30) to approximate the posterior expectation of $\log(1 + \theta^2/n)$ required in (27) yields

$$d(\tilde{\mu} \,|\, \boldsymbol{x}) \approx \frac{n}{2} \log\left[1 + \frac{1}{n} \frac{n(1+t^2) + t^4}{n + t^2 + 1}\right]. \tag{31}$$

Progressively cruder, but simpler approximations are

$$d(\tilde{\mu}\,|\,\boldsymbol{x}) \approx \frac{n}{2}\log\left[1+\frac{1}{n}\left(1+t^2\right)\right] \approx \frac{1}{2}\left(1+t^2\right). \tag{32}$$

Thus, for large n, the intrinsic expected loss $d(\tilde{\mu}\,|\,\boldsymbol{x})$ is essentially quadratic in the number $t = (\overline{x}-\tilde{\mu})/(s/\sqrt{n-1})$ of standard deviations which separate $\overline{x}$ from $\tilde{\mu}$. Summarizing, we have thus established

Theorem 2 ***(Intrinsic estimation of the Normal mean)*** *Let $\boldsymbol{x}$ be a random sample of size n from $\mathrm{N}(x\,|\,\mu,\sigma)$, with mean and variance $\overline{x}$ and s^2, and let $t = \sqrt{n-1}\,(\overline{x}-\tilde{\mu})/s)$ be the conventional t statistic.*

(i) The intrinsic point estimator of μ is $\tilde{\mu}_{int}(\boldsymbol{x}) = \overline{x}$.

(ii) The unique p-credible intrinsic region for μ is the probability centred interval

$$C_p^{int}(\boldsymbol{x}) = \overline{x} \;\pm\; q_{p,n}\; s/\sqrt{n-1}\;,$$

where $q_{p,n}$ is the $(p+1)/2$ quantile of a standard Student-t distribution with $n-1$ degrees of freedom. For all sample sizes, the frequentist coverage of $C_p^{int}(\boldsymbol{x})$ is exactly p.

(iii) The expected intrinsic loss associated to the use of $\tilde{\mu}$ as a proxy for μ is

$$d(\tilde{\mu}\,|\,\boldsymbol{x}) \approx \frac{n}{2}\log\left[1+\frac{1}{n}\,\frac{n(t^2+1)+t^4}{n+t^2+1}\right] \approx \frac{n}{2}\log\left[1+\frac{1}{n}\left(1+t^2\right)\right].$$

As a numerical illustration, a random sample of size $n = 25$ was generated from a standard normal, yielding $\overline{x} = -0.162$ and $s = 0.840$. The intrinsic estimator is $\mu^* = \overline{x} = -0.162$ and the 0.99-intrinsic credible region is the interval $[-0.642, 0.318]$. The exact value of the expected intrinsic loss $d(1/3\,|\,\boldsymbol{x})$, computed from (27) by numerical integration, is 3.768, while (31) and the two approximations in (32) respectively yield 3.781, 3.970 and 4.673. Hence, the observed data may be expected to be about $\exp(3.768) \approx 43$ times more likely under the true value of μ that under the closest normal model with $\mu = 1/3$, suggesting that the value $\mu = 1/3$ is hardly compatible with the observed data.

3.3 Intrinsic estimation of the normal variance

It has already been established (Example 2, Eq. 13) that the intrinsic discrepancy loss from using $\tilde{\sigma}^2$ as a proxy for σ^2 is

$$\delta_{\boldsymbol{x}}\{\tilde{\sigma}^2,(\mu,\sigma)\} \;=\; \delta_{\boldsymbol{x}}\{\phi\} = \begin{cases} \dfrac{n}{2}\,g(\phi) & \text{if} \quad \phi < 1, \\[2ex] \dfrac{n}{2}\,g(1/\phi) & \text{if} \quad \phi \geq 1, \end{cases} \tag{33}$$

where $g(t) = (t-1) - \log t$, and $\phi = \tilde{\sigma}^2/\sigma^2$, and that this is also the intrinsic loss $\delta_{\boldsymbol{x}}\{\tilde{\psi}, (\mu, \sigma)\}$ from using $\tilde{\psi}$ a a proxy for ψ for any one-to-one function $\psi(\sigma^2)$ of σ^2. Moreover, the reference posterior of ϕ is the gamma distribution of Eq. 18. Hence, the intrinsic estimator of σ^2 is

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) \;=\; \arg\min_{\tilde{\sigma}^2>0} \int_0^\infty \delta_{\boldsymbol{x}}(\phi)\, \mathrm{Ga}\Big(\phi \,\Big|\, \frac{n-1}{2}\,,\; \frac{n\,s^2}{2\tilde{\sigma}^2}\Big)\, d\phi,$$

where $\delta_{\boldsymbol{x}}(\phi)$ is given by (33). Moreover, it immediately follows from (18) that, as a function of σ, the reference posterior distribution of $\tau = n\,s^2/\sigma^2$ is

$$\pi(\tau \,|\, \boldsymbol{x}) = \pi(\tau \,|\, n) = \chi^2(\tau \,|\, n-1) \tag{34}$$

a central χ^2 with $n-1$ degrees of freedom; but $\phi = c\,\tau/n$, with $c = \tilde{\sigma}^2/s^2$ and, therefore, the expected posterior loss from using $\tilde{\sigma}$ may further be written as

$$d(\tilde{\sigma}^2 \,|\, s^2, n) = d(c \,|\, n) = \int_0^\infty \delta\Big(\frac{c\,\tau}{n}\Big)\, \chi^2(\tau \,|\, n-1)\, d\tau, \qquad c = \tilde{\sigma}^2/s^2. \tag{35}$$

Thus, the intrinsic estimator of the normal variance is an affine equivariant estimator of the form

$$\sigma^2_{int}(s, n) = c^*_n\, s^2, \qquad c^*_n > 0, \tag{36}$$

where c^*_n is the value of c which minimizes $d(c \,|\, n)$ in (35). The exact value of c^*_n may be numerically found by one-dimensional numerical integration, followed by numerical optimization. The first row of Table 1 displays the exact values of c^*_n for several sample sizes. However, good analytical approximations for c^*_n may be obtained.

We first consider a general approximation method. Let ω be a particular parametrization of the problem, and consider a (variance stabilizing) *reference reparametrization*

Table 1: *Exact and alternative approximate values for the intrinsic point estimator of the normal variance $\sigma^2_{int} = c^*_n s^2$.*

n	2	3	4	5	10	20	50
c^*_n	4.982	2.347	1.803	1.569	1.231	1.106	1.041
$\left(\frac{n}{n-1}\right)^2$	4.000	2.250	1.778	1.563	1.235	1.108	1.041
$\frac{n}{n-1}\, e^{1/(n-1)}$	5.437	2.473	1.861	1.605	1.242	1.110	1.041
$\frac{n}{n-2}$	—	3.000	2.000	1.667	1.250	1.111	1.042

$\phi(\omega)$ defined as one with a uniform reference prior. This is given by any solution to the differential equation $\phi'(\omega) = \pi(\omega)$, where $\pi(\omega)$ is the marginal reference prior for ω. Under regularity conditions, the sampling distribution of $\phi(\hat{\omega})$, where $\hat{\omega} = \hat{\omega}(\boldsymbol{x})$ is the MLE of ω, and the reference posterior of $\phi(\omega)$, are both asymptotically normal. Using these approximations, the intrinsic expected loss from using $\tilde{\omega}$ is found to be (Bernardo, 2005b, Theo. 4.1)

$$d(\tilde{\omega} \,|\, \boldsymbol{x}) \approx \frac{n}{2}\Big\{\sigma_\phi^2 + [\mu_\phi - \phi(\tilde{\omega})\,]^2\Big\}. \tag{37}$$

where μ_ϕ and σ_ϕ^2 are respectively the posterior mean and posterior variance of $\phi = \phi(\omega)$. This is minimized by $\phi(\tilde{\omega}) = \mu_\phi = \mathrm{E}[\phi \,|\, \boldsymbol{x}]$. Hence, in terms of any reference parametrization ϕ, the intrinsic point estimate is approximately the posterior mean μ_ϕ and, by invariance, the intrinsic estimator of any one-to-one function, $\psi = \psi(\phi)$ is approximately given by $\tilde{\psi}_{int} = \psi(\mu_\phi)$. Thus,

$$\tilde{\phi}_{int}(\boldsymbol{x}) \approx \mu_\phi = \mathrm{E}[\phi \,|\, \boldsymbol{x}], \quad \tilde{\omega}_{int}(\boldsymbol{x}) \approx \phi^{-1}\{\mu_\phi\}. \tag{38}$$

Under regularity conditions (see, e.g., Schervish, 1995, Sec. 7.1.3) the delta method may be used to obtain simple approximations to the posterior moments of ϕ is terms of those of ω, namely

$$\mu_\phi \approx \phi\{\mu_\omega\} + \sigma_\omega^2\, \phi''\{\mu_\omega\}/2\,, \tag{39}$$

$$\sigma_\phi^2 \approx \sigma_\omega^2\, [\,\phi'\{\mu_\omega\}]^2. \tag{40}$$

Substitution into (38) and (37) respectively provide useful approximations to the intrinsic point estimator of ω, and to the expected loss from using $\tilde{\omega}$ as a proxy for ω.

In the particular case of the normal variance, it is convenient to start from the parametrization in terms of the precision $\lambda = \sigma^{-2}$, whose posterior moments have simple expressions. Since reference priors are consistent under reparametrization, the reference prior for λ is $\pi(\lambda) = \pi(\sigma)|\partial\sigma/\partial\lambda| \propto \lambda^{-1}$ and, therefore, a reference parametrization is

$$\phi = \phi(\lambda) = \int^{\lambda} \pi(\lambda)\, d\lambda = \int^{\lambda} \lambda^{-1}\, d\lambda = \log \lambda.$$

Notice that the reference prior of $\phi = \log\lambda$ is indeed uniform, as it is the case for the logarithm of any other power of σ. Using (39) and (40) with the first posterior moments of λ, given in (24), yields

$$\tilde{\phi}_{int}(\boldsymbol{x}) \approx \mu_\phi = \mathrm{E}[\log\lambda \,|\, \boldsymbol{x}] \approx \log\Big(\frac{n-1}{n\,s^2}\Big) - \frac{1}{n-1}\,, \tag{41}$$

$$\sigma_\phi^2 = \mathrm{Var}[\log\lambda \,|\, \boldsymbol{x}] \approx \frac{2}{n-1}\,\cdot \tag{42}$$

By invariance, (41) directly provides an approximation to the intrinsic estimator of the variance. This has the form of a modified version of the conventional unbiased estimator $\tilde{\sigma}^2_{n-1}$; indeed, since $\sigma^2 = e^{-\phi}$,

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) = e^{-\tilde{\phi}_{int}(\boldsymbol{x})} \approx \frac{n\, s^2}{n-1}\; e^{\frac{1}{n-1}} = \tilde{\sigma}^2_{n-1}\, e^{\frac{1}{n-1}},$$

which, as shown in the third row of Table 1, provides good approximations, even for small values of n.

A better analytical approximation to the intrinsic estimator of the normal variance may be obtained making use of the particular features of this example. This is done by separately minimizing the expected value of each of the two functions which enter the definition of the intrinsic discrepancy loss $\delta_{\boldsymbol{x}}\{\phi\}$, and using the arithmetic mean of the corresponding results.

Indeed, the delta method may be used to approximate both $\mathrm{E}[g(c\tau/n)]$ and $\mathrm{E}[g(n/(c\tau))]$ in terms $\mathrm{E}[\tau\,|\,n] = n-1$ and $\mathrm{Var}[\tau\,|\,n] = 2(n-1)$. The approximation to $\mathrm{E}[g(c\tau/n)]$ is minimized by $\hat{c}^*_{1n} = n/(n-1)$, while the approximation to $\mathrm{E}[g(n/(c\tau))]$ is minimized by by $\hat{c}^*_{2n} = n(n+1)/(n-1)^2$. As one would expect, their average,

$$\hat{c}^*_n = \frac{\hat{c}^*_{1n} + \hat{c}^*_{2n}}{2} = \Big(\frac{n}{n-1}\Big)^2 = \frac{n}{n-(2-n^{-1})}\,, \tag{43}$$

provides a good approximation to the value c^*_n which minimizes (35). As shown in the second row of Table 1, the approximation remains good even for small values of n.
Combination of (36) and (43) establishes that, for all but very small n values,

$$\tilde{\sigma}^2_{int}(\boldsymbol{x}) = \tilde{\sigma}^2_{int}(s,n) \approx \Big(\frac{n}{n-1}\Big)^2\, s^2. \tag{44}$$

In view of the second expression for $\hat{c}^*_n$ in (43), a cruder approximation is given by $\tilde{\sigma}^2_{int} \approx \tilde{\sigma}^2_{n-2}$. This is larger than the MLE $\hat{\sigma}^2 = s^2$ (which divides by n the sum of squares), and also larger than the conventional unbiased estimate of the variance $\tilde{\sigma}^2_{n-1}$ (which divides by $n-1$). Notice that numerical differences between intrinsic and conventional estimators may be large for small values of n. In particular, with only two observations $\{x_1, x_2\}$, the intrinsic estimator of the variance is $\sigma^2_{int}(2, s^2) \approx 5\, s^2 = 5(x_1 - x_2)^2/4$; this is 2.5 times larger than the unbiased estimator, $(x_1 - x_2)^2/2$ in this case, which (with good reason) is generally considered to be too small.

As shown by (35), the expected intrinsic loss $d(\tilde{\sigma}^2\,|\,n, s^2)$ of any affine equivariant estimator of the variance $\tilde{\sigma}^2 = k_n s^2$, is actually independent of s^2 and only depends on the sample size n. Moreover, it is easily verified that the expected intrinsic loss $d(\tilde{\sigma}^2\,|\,n, s^2)$ is precisely equal to the *frequentist risk* associated to the intrinsic discrepancy loss,

$$r(\tilde{\sigma}^2_i\,|\,n, \sigma^2) = \int_0^\infty \delta\{\tilde{\sigma}^2_i(s^2), \sigma^2\}\, p(s^2\,|\,n, \sigma^2)\, ds^2.$$

Thus, under intrinsic discrepancy loss, the intrinsic estimator $\tilde{\sigma}^2_{int}$ dominates all affine equivariant estimators. For details, see Bernardo (2006).

Region estimation is now considered. As described in Example 4, the intrinsic p-credible region for σ is the unique solution $C_p^{int} = \{\sigma_0, \sigma_1\}$ to the equations system

$$\Big\{ d(\sigma_0^2 \,|\, \boldsymbol{x}) = d(\sigma_1^2 \,|\, \boldsymbol{x}), \quad \int_{\sigma_0}^{\sigma_1} \pi(\sigma \,|\, \boldsymbol{x})\, d\sigma = p \Big\}.$$

Using (34), this may equivalently be written in terms of $\tau = n\, s^2/\sigma^2$ as

$$\Big\{ d(\sigma_0^2 \,|\, \boldsymbol{x}) = d(\sigma_1^2 \,|\, \boldsymbol{x}), \quad \int_{ns^2/\sigma_1}^{ns^2/\sigma_0^2} \chi(\tau \,|\, n-1)\, d\tau = p \Big\}. \tag{45}$$

Thus, the unique p-credible intrinsic region for σ^2 is the interval

$$C_p^{int}(\boldsymbol{x}) = \Big\{ \frac{n\, s^2}{\chi^2_{n-1}(1-\alpha)} \,,\, \frac{n\, s^2}{\chi^2_{n-1}(1-p-\alpha)} \Big\} \tag{46}$$

where $\chi^2_{n-1}(q)$ is the q quantile of a χ^2_{n-1} distribution, and α is the solution to the equation

$$d\Big(\frac{n\, s^2}{\chi^2_{n-1}(1-\alpha)} \,|\, \boldsymbol{x} \Big) = d\Big(\frac{n\, s^2}{\chi^2_{n-1}(1-p-\alpha)} \,|\, \boldsymbol{x} \Big). \tag{47}$$

By invariance, this provides the intrinsic p-credible region of any one-to-one function of σ^2.

As a function of the data $\boldsymbol{x}$, the sampling distribution of $n\, s^2/\sigma^2$ is also a χ^2 with $n-1$ degrees of freedom. Hence, for all sample sizes, the expected coverage under sampling of the p-credible intervals (46) is *exactly* p.

Using (35) to evaluate expected losses, the exact solution to equation (47) may easily be obtained by numerical methods. However, good analytical approximations are may be obtained.

Working again in terms of the reference parametrization for this problem, $\phi = \log \lambda = -2 \log \sigma$, and using (37), (39) and (40), the expected loss from using $\tilde{\phi}$ as a proxy for ϕ is approximately

$$d(\tilde{\phi} \,|\, \boldsymbol{x}) \approx \frac{n}{2} \left[\frac{2}{n-1} + (\, \tilde{\phi}_{int} - \tilde{\phi}\,)^2 \right]. \tag{48}$$

But this is symmetric around $\tilde{\phi}_{int} = \log(\tilde{\lambda}_{int}) = -\log(\sigma^2_{int})$ and therefore, to keep those $\tilde{\phi}$ points with smaller expected loss, any intrinsic credible region for $\phi = \log \lambda$ must be (approximately) centered at $\tilde{\phi}_{int}$. Thus, using (42) and (44) this will be of the form

$$C_p^{int}(\boldsymbol{x}, \Phi) \approx \tilde{\phi}_{int} \pm \alpha_{pn}\sigma_\phi \approx \log\Big[\Big(\frac{n-1}{n}\Big)^2 \frac{1}{s^2}\Big] \pm \gamma_{pn}\sqrt{\frac{2}{n-1}} \tag{49}$$

where γ_{pn} is the solution to the equation

$$\int_{\tilde{\phi}_{int}-\gamma_{pn}\,\sigma_\phi}^{\tilde{\phi}_{int}+\gamma_{pn}\,\sigma_\phi} \pi(\phi\,|\,\boldsymbol{x})\,d\phi = p,$$

or, equivalently since

$$\begin{aligned} \tau = n\,s^2\lambda &= n\,s^2 e^{\phi}, \\ ns^2\,e^{\tilde{\phi}_{int}\pm\gamma_{pn}\,\sigma_\phi} &= \frac{(n-1)^2}{n}\,\exp\Big[\,\pm\gamma_{pn}\sqrt{\frac{2}{n-1}}\,\Big], \\ \pi(\tau\,|\,\boldsymbol{x}) &= \chi^2(\tau\,|\,n-1), \end{aligned}$$

γ_{pn} is the unique solution to the equation

$$F_{n-1}\Big\{\frac{(n-1)^2}{n}\;e^{+\gamma_{pn}\sqrt{\frac{2}{n-1}}}\Big\} - F_{n-1}\Big\{\frac{(n-1)^2}{n}\;e^{-\gamma_{pn}\sqrt{\frac{2}{n-1}}}\Big\} = p, \tag{50}$$

where F_ν is the cumulative distribution function of a χ^2_ν distribution.

A numerical solution to (50) is immediately found with standard statistical software. However, a simple analytical approximation may be derived using the fact that the reference posterior distribution of $\phi = \log\lambda$ becomes approximately normal (at a faster rate that any other simple function of σ) as the sample size n increases. Using this approximation and (49), the p-credible intrinsic region for ϕ is approximated by the interval

$$C_p^{int}(\boldsymbol{x}, \Phi) \approx \log\Big[\Big(\frac{n-1}{n}\Big)^2 \frac{1}{s^2}\Big] \pm q_p\sqrt{\frac{2}{n-1}} \tag{51}$$

where q_p is the $(p+1)/2$ quantile of a standard normal distribution. By invariance, the p-credible intrinsic region for the variance $\sigma^2 = e^{-\phi}$ will be approximated by

$$\Big\{\; s^2\Big(\frac{n}{n-1}\Big)^2\; e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}},\;\; s^2\Big(\frac{n}{n-1}\Big)^2\; e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\;\Big\} \tag{52}$$

where γ_{np} is the solution to (50) which, as n increases, converges to q_p, the $(p+1)/2$ quantile of an standard normal distribution.

Summarizing, we have thus established

Theorem 3 ***(Intrinsic estimation of the normal variance)*** *Let* $\boldsymbol{x}$ *be a random sample of size* n *from* $\mathrm{N}(x\,|\,\mu,\sigma)$*, with variance* s^2.

(i) The intrinsic point estimator $\tilde{\sigma}^2_{int}(\boldsymbol{x})$ of σ^2 is

$$\begin{aligned}
\tilde{\sigma}^2_{int}(\boldsymbol{x}) &= \arg\min_{\tilde{\sigma}>0} d(\tilde{\sigma}^2 \,|\, \boldsymbol{x}),\\
d(\tilde{\sigma}^2 \,|\, \boldsymbol{x}) &= \frac{n}{2}\int_0^\infty \delta\left(\frac{\tilde{\sigma}^2\,\tau}{n\,s^2}\right)\chi^2(\tau \,|\, n-1)\,d\tau,\\
\delta\{\theta\} &= \min\{g(\theta), g(1/\theta)\}, \quad g(\theta) = (\theta-1) - \log\theta,\\
\tilde{\sigma}^2_{int}(\boldsymbol{x}) &\approx \left(\frac{n}{n-1}\right)^2 s^2.
\end{aligned}$$

The intrinsic point estimator $\tilde{\sigma}_{int}(\boldsymbol{x})$ is the Bayes estimator with respect to the intrinsic discrepancy loss. Besides, it has smaller frequentist risk with respect to this loss than any other affine equivariant estimator.

(ii) The unique p-credible intrinsic region $C^{int}_p(\boldsymbol{x})$ for σ^2 is the interval

$$C^{int}_p(\boldsymbol{x}) = \{a(\alpha,\boldsymbol{x}),\ b(\alpha,p,\boldsymbol{x})\} = \left\{\frac{n\,s^2}{\chi^2_{n-1}(1-\alpha)}\ ,\ \frac{n\,s^2}{\chi^2_{n-1}(1-p-\alpha)}\right\},$$

where $\chi^2_{n-1}(q)$ is the q quantile of a χ^2_{n-1} distribution, and α is the solution to the equation $d\{a(\alpha,\boldsymbol{x}) \,|\, \boldsymbol{x})\} = d\{b(\alpha,p,\boldsymbol{x}) \,|\, \boldsymbol{x})\}$. For all sample sizes, the frequentist coverage of $C^{int}_p(\boldsymbol{x})$ is exactly p. Moreover,

$$C^{int}_p(\boldsymbol{x}) \approx s^2\left(\frac{n}{n-1}\right)^2\left\{e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}},\ e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\}$$

where γ_{np} is the solution to the equation

$$F_{n-1}\left\{\frac{(n-1)^2}{n}\,e^{+\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\} - F_{n-1}\left\{\frac{(n-1)^2}{n}\,e^{-\gamma_{np}\sqrt{\frac{2}{n-1}}}\right\} = p.$$

and F_ν is the cumulative distribution function of a χ^2_ν distribution. As n increases, γ_{np} converges to the $(p+1)/2$ normal quantile.

(iii) The expected intrinsic loss associated to the use of $\tilde{\sigma}^2$ is

$$d(\tilde{\sigma}^2 \,|\, s^2, n) \approx \frac{n}{2}\left[\frac{2}{n-1} + \left(\log\frac{1}{\sigma^2_{int}} - \log\frac{1}{\tilde{\sigma}^2}\right)^2\right],$$

with $\tilde{\sigma}^2_{int}(\boldsymbol{x}) \approx n^2\,s^2/(n-1)^2$.

For the numerical illustration considered in Example 4 (where the sample size was only $n = 12$), the approximation (44) to the intrinsic estimate of σ^2 yields $\tilde{\sigma}^2_{int} \approx 5.104$. The approximation (52) to the intrinsic 0.90-credible region yields $(2.480, 11.507)$ using the exact solution $\gamma_{np} = 1.693$ to equation (50), and $(2.531, 11.272)$ using the

corresponding normal approximation $\gamma_{np} \approx q_{0.90} = 1.645$. These approximations may be compared eith the exact values $\tilde{\sigma}^2_{int} = 5.090$ and $C^{int}_{0.90} = (2.481, 11.717)$ numerically found in Example 4.

3.4 Intrinsic estimation of the Cauchy scale parameter

We finally consider an example where no analytical expressions are possible. With the use of increasingly complex statistical models, this is fast becoming the norm, rather than the exception, in statistical practice.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be a random sample from a Cauchy distribution $\mathrm{Ca}(x \,|\, 0, \sigma)$, centered at zero with unknown scale parameter σ, so that the likelihood function is

$$p(\boldsymbol{x} \,|\, \sigma) = \prod_{i=1}^{n} \mathrm{Ca}(x_j \,|\, 0, \sigma) \propto \sigma^{-n} \prod_{i=1}^{n} \Big(1 + \frac{x_j^2}{\sigma^2}\Big)^{-1}.$$

The Cauchy distribution does not belong to the exponential family and, therefore, there is no sufficient statistic of finite dimension. There is no analytical expression for the MLE $\hat{\sigma}$ of the unknown parameter. Fisher information function is $n/(2\sigma^2)$ and, therefore, the posterior distribution of σ will be asymptotically normal, $\mathrm{N}(\sigma \,|\, \hat{\sigma},\ \sqrt{2}\,\hat{\sigma}/\sqrt{n})$.

Since this is a scale model, the reference prior is $\pi(\sigma) = \sigma^{-1}$ and, using Bayes theorem, the reference posterior is

$$\pi(\sigma \,|\, \boldsymbol{x}) = \frac{\sigma^{-1} p(\boldsymbol{x} \,|\, \sigma)}{\int_0^\infty \sigma^{-1} p(\boldsymbol{x} \,|\, \sigma)\, d\sigma}\,, \tag{53}$$

which may easily be numerically computed. It may be verified that, provided the data $\boldsymbol{x}$ contain at least two different observations, $\pi(\sigma \,|\, \boldsymbol{x})$ has a gamma-like shape with a unique mode.

Figure 4 represents the reference posteriors of σ which correspond to a set of 25 random samples of size $n = 12$, which were all generated from a Cauchy distribution $\mathrm{Ca}(x \,|\, 0, 2)$. This may be seen as a graphical representation of the *sampling distribution of* $\pi_\sigma(\cdot \,|\, \boldsymbol{x})$, the reference posterior of σ, given $\sigma = 2$ and $n = 12$. Notice that, although all these posteriors contain indeed the true value $\sigma = 2$ from which the samples have been simulated (marked in the figure with a solid dot), the variability is very large.

The logarithmic divergence $\kappa\{\sigma_2 \,|\, \sigma_1\}$ of $\mathrm{Ca}(x \,|\, 0, \sigma_2)$ from $\mathrm{Ca}(x \,|\, 0, \sigma_1)$ is

$$\int_{-\infty}^{\infty} \mathrm{Ca}(x \,|\, 0, \sigma_1) \log \frac{\mathrm{Ca}(x \,|\, 0, \sigma_1)}{\mathrm{Ca}(x \,|\, 0, \sigma_2)}\, dx = \log \frac{1}{4\,\sigma_1\,\sigma_2} + 2 \log(\sigma_1 + \sigma_2).$$

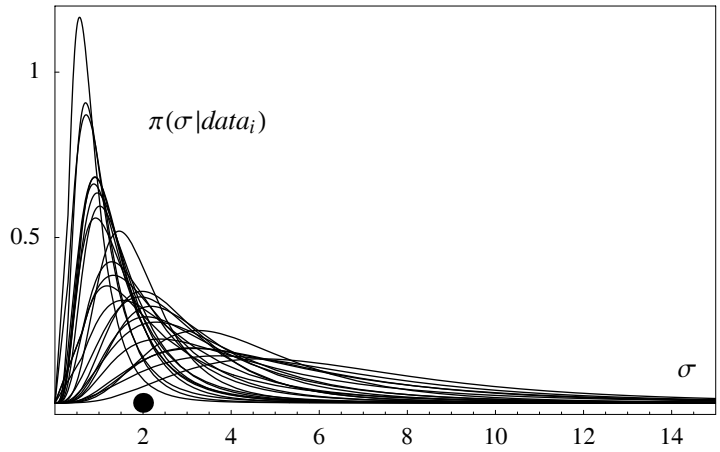

Figure 4: *Reference posterior distributions of σ for a set of 25 random samples of size $n = 5$ generated from a $\mathrm{Ca}(x\,|\,0, 2)$ distribution.*

Since, in this case, $\kappa\{\sigma_2\,|\,\sigma_1\} = \kappa\{\sigma_1\,|\,\sigma_2\}$, the intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy fo σ is (Def. 2)

$$\delta\{\tilde{\sigma}, \sigma\} = \log \frac{1}{4\,\tilde{\sigma}\,\sigma} + 2\log(\tilde{\sigma} + \sigma) = \log \frac{1}{4\,\sqrt{\phi}} + \log(1 + \sqrt{\phi}), \tag{54}$$

where $\phi = \tilde{\sigma}^2/\sigma^2$. Thus, as in the normal case (Eq. 33), the intrinsic discrepancy loss only depends on the variance ratio $\phi = \tilde{\sigma}^2/\sigma^2$.

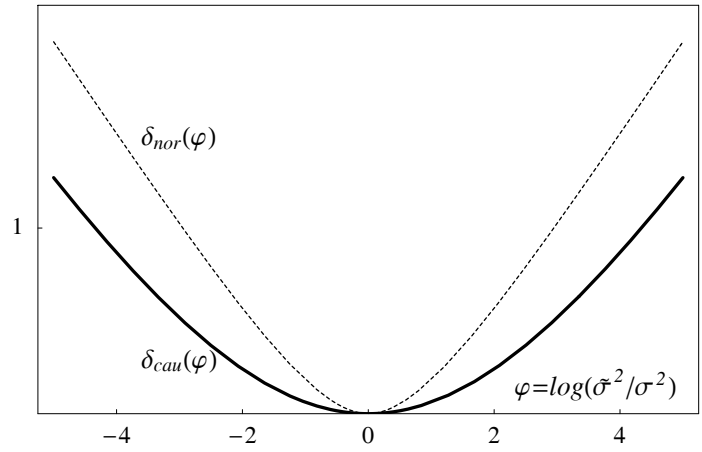

Figure 5: *Intrinsic discrepancy loss from using $\tilde{\sigma}$ as a proxy for σ as a function of $\psi = \log(\tilde{\sigma}^2/\sigma^2)$ for Cauchy (solid line) and normal (dotted line) distributions.*

Figure 5 provides a direct comparison between the intrinsic discrepancy loss for the scale parameter in the Cauchy and in the normal case. As one might expect, for any

given value of the ratio $\phi = \tilde{\sigma}^2/\sigma^2$, the intrinsic loss is smaller in the Cauchy case than it is in the normal case.

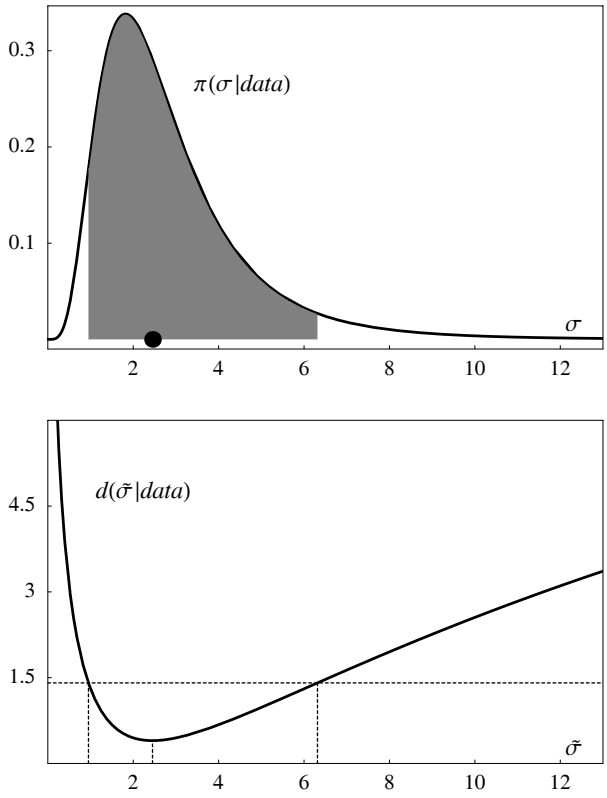

Figure 6: *Reference posterior density (upper panel) and intrinsic expected loss (lower panel) for the scale parameter σ of a Cauchy* $\mathrm{Ca}(x\,|\,0,\sigma)$ *distribution, given* $\boldsymbol{x} = \{-1.78, -0.75, -2.44, -3.30, 8.48\}$. *The intrinsic estimator is* $\tilde{\sigma}_{int} = 2.452$ *(solid dot) and the intrinsic* 0.90*-credible region is* $C^{int}_{0.90} = (0.952, 6.314)$ *(shaded region).*

The intrinsic expected loss from using $\tilde{\sigma}$ is the reference posterior expectation of the intrinsic discrepancy loss,

$$d(\tilde{\sigma}\,|\,\boldsymbol{x}) = n \int_0^\infty \delta\{\tilde{\sigma}, \sigma\}\, \pi(\sigma\,|\,\boldsymbol{x})\, d\sigma,$$

where $\delta\{\tilde{\sigma}, \sigma\}$ and $\pi(\sigma\,|\,\boldsymbol{x})$ are respectively given by (54) and (53), and may easily be computed by numerical integration. The intrinsic estimator of σ is

$$\tilde{\sigma}_{int}(\boldsymbol{x}) = \arg\inf_{\tilde{\sigma}>0} d(\tilde{\sigma}\,|\,\boldsymbol{x})$$

and the p-credible intrinsic region is the solution $C_p^{int}(\boldsymbol{x}) = (\sigma_0, \sigma_1)$ to the equations system

$$\left\{ d(\sigma_0 \,|\, \boldsymbol{x}) = d(\sigma_1 \,|\, \boldsymbol{x}), \quad \int_{\sigma_0}^{\sigma_1} \pi(\sigma \,|\, \boldsymbol{x})\, d\sigma = p \right\}.$$

As a numerical illustration, a random sample of size $n = 5$ was generated from a Cauchy $\mathrm{Ca}(x \,|\, 0, 2)$, yielding $\boldsymbol{x} = \{-1.78, -0.75, -2.44, -3.30, 8.48\}$. The results from the reference analysis of this data set are encapsulated in Figure 6. The reference posterior distribution $\pi(\sigma \,|\, \boldsymbol{x})$ is represented in the upper panel, and the expected intrinsic loss $d(\tilde{\sigma} \,|\, \boldsymbol{x})$ from using $\tilde{\sigma}$ as a proxy for σ is represented in the lower panel. The intrinsic estimator, represented by a solid dot, is $\tilde{\sigma}_{int} = 2.452$, and the intrinsic 0.90-credible interval, represented by a shaded region, is $C_{0.90}^{int} = (0.952, 6.314)$.

Neither exact Bayesian credible regions nor exact frequentist confidence intervals may be analytically obtained in this problem. The frequentist coverage of the intrinsic credible regions was however analyzed by simulation. A set of $5,000$ random samples of size $n = 5$ were generated from a Cauchy $\mathrm{Ca}(x \,|\, 0, 2)$, and their corresponding intrinsic 0.90-credible intervals were computed; it was then found that the proportion of those intervals which contained the true value $\sigma = 2$ was 0.905. With $25,000$ random samples this proportion was 0.902. This suggests that (as in the normal case) the expected frequentist coverage of reference p-credible regions, the limit of this algorithm as the number of generated random samples increases, is exactly p. To further explore this suggestion, a set of $10,000$ random samples of size n were generated from a Cauchy distribution $\mathrm{Ca}(x \,|\, , 0, \sigma)$ for each of several combinations $\{n, \sigma\}$ of sample size n and true value σ of the scale parameter and the corresponding intrinsic p-credible regions where computed for $p = 0.90$ and $p = 0.95$. Table 2 describes the proportion of these regions which actually contained the value of σ from which the samples had been generated.

Table 2: *Proportions of intrinsic p-credible intervals which contained the true value of σ among $10,000$ random samples generated from each of several combinations of sample size n and true value of σ.*

	$p = 0.90$			$p = 0.95$		
	n			n		
σ	2	12	30	2	12	30
0.5	0.9002	0.9044	0.8999	0.9490	0.9491	0.9507
2.0	0.8971	0.8971	0.9003	0.9467	0.9517	0.9490
4.0	0.9006	0.8960	0.8990	0.9484	0.9497	0.9507

Examination of this table provides strong statistical evidence that the frequentist coverage of reference p-credible regions is indeed exactly equal to p for all sample sizes. Indeed, treating each simulation as a Bernoulli trial, the reference posterior

distribution of the frequentist coverage θ_{ij} which corresponds to the (i, j) cell is approximately normal with mean observed proportion p_{ij} quoted in the table, and standard deviation $(0.90 * 0.10/10000)^{1/2} = 0.0030$ for the 0.90-credible intervals, and $(0.95 * 0.05/10000)^{1/2} = 0.0022$ for the 0.95-credible intervals. This makes the respective nominal values 0.90 and 0.95 clearly compatible with the observed results. Notice that this is *not* an asymptotic analysis, as in probability matching theory (Datta and Sweeting, 2005), for it even applies to the smallest possible samples, those with $n = 2$.

References

Berger, J. O. and Bernardo, J. M. (1992). On the development of reference priors. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 35-60 (with discussion).

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical*, 41, 113-147 (with discussion). Reprinted in *Bayesian Inference* (N. G. Polson and G. C. Tiao, eds.) Brookfield, VT: Edward Elgar, 1995, 229-263.

Bernardo, J. M. (2005a). Reference analysis. *Handbook of Statistics* 25 (D. K. Dey and C. R. Rao, eds.). Amsterdam: Elsevier, 17-90.

Bernardo, J. M. (2005b). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, 14, 317-384 (with discussion).

Bernardo, J. M. (2006). Intrinsic point estimation of the normal variance. *Bayesian Statistics and its Applications* (S. K. Upadhyay, U. Singh and D. K. Dey, eds.) New Delhi: Anamaya, 110-121.

Bernardo, J. M. and Juárez, M. (2003). Intrinsic estimation. *Bayesian Statistics*, 7, 465-476.

Bernardo, J. M. and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. *International Statistical Review*, 70, 351-372.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley. (2nd edition in preparation).

Brown, L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *The Annals of Mathematical Statistics* ,39, 24-48.

Brewster, J. F. and Zidek, J. V. (1974). Improving on equivariant estimators. *Annals of Statistics*, 2, 21-38.

Brown, L. (1990). Comment on Maata and Casella (1990).

Datta, G. S. and Sweeting, T. J. (2005). Probability matching priors. *Handbook of Statistics*, 25 (D. K. Dey and C. R. Rao, eds.). Amsterdam: Elsevier, 91-114.

Fernández, C. and Steel, M. F. J. (1999). Reference priors for the general location-scale model. *Statistics and Probability Letters,* 43, 377-384.

Juárez, M. A. (2004). *Métodos Bayesianos Objetivos de Estimación y Contraste de Hipótesis*. Ph.D. Thesis, Universidad de Valencia, Spain.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.

Maata, J. M. and Casella, G. (1990). Developments in decision-theoretic variance estimation. *Statistical Science*, 5,90-120, (with discussion).

Robert, C. P. (1996). Intrinsic loss functions. *Theory and Decision,* 40, 192-214.

Rukhin, A. L. (1987). How much better are better estimators of a normal variance. *Journal of The American Statistical Association*, 82, 925-928.

Schervish, M. J. (1995). *Theory of Statistics*. Berlin: Springer

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, 16, 155-160.

Discussion of "Objective bayesian point and region estimation in location-scale models" by José M. Bernardo

Miguel A. Gómez Villegas

Departamento de Estadística e Investigación Operativa

Universidad Complutense de Madrid

Let me begin by congratulating Professor Bernardo for his excellent job in objective Bayesian analysis. This paper, and the closely related Bernardo (2005), present a unified theory of estimation by point and credible regions based on information ideas he has used previously to define reference priors. The idea originates from the study of both problems as decision problems, where the loss function is the "intrinsic discrepancy" inspired in the Kullblack-Leibler divergence, and defined as the minimum of $k_x\{\widetilde{\theta}, \widetilde{\lambda}|\theta, \lambda\}$ and $k_x\{\theta, \lambda|\widetilde{\theta}, \widetilde{\lambda}\}$ where

$$k_x\{\widetilde{\theta}, \widetilde{\lambda}|\theta, \lambda\} = \int_{\chi(\theta,\lambda)} \pi(x|\theta, \lambda)\ ln\frac{\pi(x|\theta, \lambda)}{\pi(x|\widetilde{\theta}, \widetilde{\lambda})}dx$$

An intrinsic point estimator is then defined as the Bayes estimator which corresponds to the intrinsic loss and the appropriate reference prior. A p-credible intrinsic region estimator is defined as the lowest posterior loss p-credible with respect to the intrinsic loss and the appropriate reference prior.

A first question is: do we need to employ

$$\int_{C_p^{int}} \pi(\theta|x)d\theta \geq p$$

with the inequality instead of equality to allow the discrete case?

Second, it would be useful to have a better understanding of the proposed approach to applying these ideas to the exponential distribution family instead of location-scale models; this is a family of distributions greater than the other.

Professor Bernardo claims that in one-dimensional problems, one may define probability centred credible intervals, and these are invariant under reparametrization. Will it not be necessary to suppose that the transformation is monotonic?

Third, on a more philosophical basis, I think that invariance is a compelling argument for point estimations and for credible regions. Indeed both point estimations and credible regions are two answers to the same question: how we can eliminate the uncertainty

about θ. Bernardo's approach permits one to obtain invariance under reparametrization in both problems.

Fourth, the chosen examples show the coherence between frequentist inference and Bayesian inference. When intrinsic credible regions that require minimal subjective inputs are employed, exact frequentist confidence regions are obtained, at least in the normal mean and variance. This fact is similar to the one obtained by this discussant in Gómez-Villegas and González-Pérez (2005) and references therein. I wonder if Professor Bernardo has any idea about the essential reasons behind the matching properties between intrinsic credible regions and confidence regions in these cases?

Fifth, adopting this approach to credible set construction, I see problems in computations, the posterior intrinsic loss integrated over a large dimensional space. From the point of view of applications, a simple asymptotic approximation to normality should be necessary.

In closing, I would like to thank the editor of the journal for giving me the opportunity to discuss this paper.

References

Bernardo, J. M. (2005). Intrinsic credible regions: an objective Bayesian approach to interval estimation. *Test*, 14, 317-384.

Gómez-Villegas, M. A. and González-Pérez, B. (2005). Bayesian analysis of contingency tables. *Communications in Statistics-Theory and Methods*, 34, 1743-1754.

Dennis V. Lindley

ThomBayes@aol.com

Two concepts are basic to the ideas of this excellent paper: *objectivity* and the concept of estimation as a *decision* problem. In the author's skilful hands, these lead to reference priors and intrinsic loss functions, and hence, by minimizing expected loss, to estimates which are often superior to the conventional ones. It can be said with some confidence that we have here a solution to the problem Harold Jeffreys first posed around 1939 of providing an objective, coherent method for scientific inference. The development employs several subtle ideas, and considerable mathematical complexity, but one feature that struck me is that the final results are usually fairly simple and look right. An example of this is provided by the loss functions in Figure 3, which have the reasonable convexity property around the true value but, unlike quadratic loss, exhibit sensible concavity at more discrepant values. I would have preferred the loss to have been bounded but, with normal distributions and their thin tails, this scarcely matters. To be bounded may be more important with the fat tails of the Cauchy in Figure 6, in order to avoid paradoxes of the St. Petersburg type. A related point is that although the mathematics can be formidable, at least in the view of some applied statisticians, once it has been done the practitioner can easily use the results in the confidence that the machinery used to produce them is sound. It is comparable to driving a car, without knowing how it was made, but having confidence in the manufacturer.

Granted the basic concepts, this is an important paper, but was Jeffreys right to search for objectivity, and was Fisher wrong in dismissing decision concepts from inference? I think Jeffreys was wrong and Fisher was right. At the risk of repeating what I have said before, it seems to me that inference and decision-making are distinct and both are subjective. In other words, the two basic concepts, that provide the foundations of this paper, are suspect.

Consider first the fixed likelihood upon which all the arguments in the paper rest. Is it really objective? There are a few cases where substantial evidence for normality exists, but often the normal, or another member of the exponential family, is used merely for mathematical simplicity. With the increased computing power available today, statisticians are less constrained and can use other distributions that appear more realistic, thereby introducing subjectivity. There are some popular data sets that have been repeatedly analysed using different likelihoods. Where is the objectivity there? It is interesting that Bernardo uses one symbol, p, for probabilities of data but another,

π, for probabilities of parameters. In reality both p and π reflect beliefs about data and parameters respectively, obey the same rules and do not deserve separate treatments.

Inference concerns parameters. (It is more practical to make inference about future data but I do not explore that trail here.) What are these parameters, the θ and λ of the paper? If our statistical analyses are to be of use in data analysis, θ at least ought to relate to something in the real world. Bernardo has only one sentence about this, referring to θ as the age of the earth. Putting aside intelligent designers, reputable scientists differ in their views of the age. In other words, their ideas are subjective, so that before relevant data about the age are considered, their different views need to be included. Another relevant fact is that information about the age of the earth does not come from data with normal, or any other objective, likelihood. More conspicuous examples of subjectivity are apparent witch clinical trials, where the different views of drug companies and official bodies are consulted before the trial. This became clear recently when a trial went horribly wrong and experts claimed the probabilities used were, in their opinion, unsound. So often today, θ is regarded as nothing more than a symbol, whereas, to be of value, it has to refer to reality and hence influenced by opinions about that reality. These opinions should be incorporated into the analysis, not ignored and replaced by a reference prior, especially when this is improper.

All of us have, at some time, expressed an opinion about something without having any intention of basing any action upon that opinion. In statistics, this opinion-forming is inference and means we infer the value of the real θ. In the Bayesian paradigm this is done by means of your probability distribution of θ, given the data and the original information about θ. Whilst it is true that any inference has to be capable of being used as a basis for action, for otherwise what use is it, it is not true that inference has to have immediate actions in mind. In particular, inference does not require a loss function, and certainly not a loss function that ignores reality. In Bayesian terms, there is only one inference, the posterior distribution and, although it may be advantageous to summarize its main features, such approximations scarcely need elaborate techniques, except in the case of many parameters.

Inference from data consists in modelling that data in the form of a likelihood depending on parameters, supplying your opinion of the parameters prior to the data, and combining likelihood and prior by Bayes theorem. Finally the nuisance aspect of the parameters is removed by integration. When several people are involved there may be disagreements over likelihood or prior. These may be removed by discussion but, if this fails, the calculations may be repeated under different subjective opinions and the posteriors compared. That science is objective is a myth. Apparent objectivity in science only arises when the data are extensive.

This paper explores a field that, in my view, is not in the broad stream of statistics. This is not to deny it great merit, for we now know what that field contains, material of real merit from which all can learn.

Mark J. Schervish

Carnegie Mellon University, USA

I admire Professor Bernardo for his steadfastness and resolution in staying the course of research into reference priors and other so-called objective Bayesian methods. Despite repeated attacks dating back to the discussion of Bernardo (1979) he has continually risen to the challenge of making these methods palatable to practitioners and theoreticians alike. I will not here rehearse all of the criticisms or the support for his work in this area. I refer the interested reader to the various discussions of the papers listed in the reference list to Professor Bernardo's paper. I will mention just a few problems that I have with the methods as well as what I like about them.

To begin with a positive note, I like the idea of having a transformation-equivariant estimation procedure for non-decision-theoretic inference. When one is faced with a decision problem in which a specific loss function is relevant, then one does not care whether one's inference satisfies an ad hoc criterion such as transformation equivariance. On the other hand, when one merely wishes to report an estimate of some quantity, especially the parameter of a statistical model which most likely is a figment of one's imagination (model) anyway, then it becomes difficult to explain why the estimate of an equivalent parameter is not the equivalent estimate. Indeed, I believe that the intrinsic discrepancy loss satisfies a slightly stronger invariance than is stated in (10). I believe that one could apply a one-to-one reparameterization of the form $\phi = \phi(\theta)$ and $\psi = \psi(\lambda, \theta)$ and still achieve (10). Of course, a completely general reparameterization would change the meaning of the parameter of interest, and yet the desire for an equivariant estimate would remain.

One of the serious concerns with reference priors is their violation of the likelihood principle. The reference priors are different for binomial sampling and negative binomial sampling so that even if the observed data could have come from either sampling scheme, the posterior would depend on the sampling plan. If one were to observe a binomial sample and use the reference prior, and later observe a negative binomial sample, one would get a different inference than if one were to observe the same two samples in the other order. As mentioned earlier, various discussants have described other concerns with the methods advocated in the manuscript, and I will let the reader find them in their original forms. I will add only one other concern that I have, and that is with the use of the description of these methods as "objective". I suppose that, so long as one agrees with all of the reasons put forth for why such methods should be

used, then one will use the methods and they become objective in that sense. But any set of methods could be called objective on those grounds. One of the main strengths of Bayesian methodology is that it forces users to be explicit about the assumptions that they are making. People who think that they are using objective methods are simply borrowing a collection of subjective assumptions and ignoring the fact that choices were made by someone else arriving at those assumptions. When you lay your assumptions out for all to see, you are in a position to evaluate the sensitivity of your inferences to the assumptions. If you hide behind a cloak of objectivity, you may produce the same answer that others produce, but you have lost the ability to see what is the effect of the subjective choices that were made.

Rejoinder

I am extremely grateful to the three discussants by their thoughtful comments. I will answer them individually.

Gómez-Villegas. If the parameter of interest $\boldsymbol{\theta}$ is discrete, then we would certainly need to work with regions C such that $\int_C \pi(\theta \mid \boldsymbol{\theta})\, d\boldsymbol{\theta} \geq p$ since, in that special case, not all credible probabilities p would be attainable. However, point and region estimation are usually done with *continuous* parameter spaces, and this is indeed the case in the location and scale models considered in this paper. In that situation, the equality may always be obtained.

The ideas discussed in the paper may certainly be applied to models in the (generalized) exponential family and it is likely that this would lead to some rather general results. I did not have time and space to do this here, but it is certainly a research line well worth exploring.

As Professor Gómez-Villegas points out, the invariance arguments invoked only refer to monotonic, one-to-one transformations of the parameter. Even though not always explicitly stated, we were indeed always assuming this to be the case.

I believe that the *exact* numerical coincidence between objective credible regions and frequentist confidence interval is the exception, not the rule; when it happens, it is the consequence of the existence of pivotal quantities, so that the reference distribution of the pivot (considered as a function of the parameter) is precisely the same as its sampling distribution (considered as a function of the data). In particular, this coincidence cannot exist if data are discrete, as in the case of binomial or Poisson data. Beyond the particular situations where pivots exist, one may only expect an asymptotic approximation: objective credible regions are typically *approximate* confidence intervals, the approximation improving with the sample size.

Routine application of the methods described in this paper will certainly require either available software producing the exact results (not difficult to write in the standard examples which constitute the vast majority of applications) and/or appropriate analytical approximations. The latter may easily be obtained, as in the examples contained in the paper, by using the normal approximation with the parametrization induced by the appropriate variance-stabilizing transformation, and then making use of the invariance properties of the procedures.

Lindley. I am really proud that Professor Lindley may believe that the procedures described provide an objective coherent method for scientific inference in the sense demanded by Jeffreys, and I am very grateful for that comment.

It would certainly be better from a foundations viewpoint if the expected loss were bounded, but information measures with continuous parameters are are not bounded (one needs infinite amount of information to know precisely a real number) and yet have all kind of attractive properties.

To repeat in print the basics of an argument that Professor Lindley and I have often had in private conversations,

(i) I believe, with Jeffreys, that Fisher was wrong in dismissing decision concepts in inference. If, by some reason, you must choose an estimate, then (whether you like it or not) you have a well posed decision problem where the action space is the set of parameter values; then foundations dictate that (to act rationally) you *must* use a loss function. For instance, in one continuous parameter problems, the median may well be an estimate with good robustness properties, but the fact remains that this would be a good estimate if (*and only if*) your loss function is well approximated by a linear, symmetric loss function.

(ii) I applaud the use of subjective priors when the problem is simple and small enough for the required probability assessments to be feasible (which is *not* frequent). But, even in this case, there is no reason while other people should necessarily accept a subjective prior which goes beyond clearly stated assumptions and verifiable (possibly historical) data. There is a clear need for some commonly accepted minimum set of conclusions to be solely derived from assumptions and data, and this is precisely what reference posteriors provide. As their name indicate, they are proposed as a *reference*, to be compared with subjective posteriors when these are available. This is part of a necessary exercise in sensitivity analysis, by making explicit which parts of the conclusions depend on a particular subjective prior, and which parts are implied by the model assumed and the data obtained.

As Professor Lindley points out, although inferential statements are typically used as a basis for action, there are many situations were inferences are to be drawn without any specific action in mind. This is precisely why we suggest the use of an the information-based loss function. If a particular action is in mind, one should certainly use a context dependent loss function which appropriately describes the decision problem analyzed. It no particular decision problem is in mind, one is bound to use some conventional loss function. We have argued that conventional loss functions (such as the ubiquitous quadratic loss) are often unsatisfactory. Instead, for "pure inference" problems one should try to minimize the information loss due to the use of an estimate of the unknown parameter value; and this, I believe, is appropriately captured by the intrinsic discrepancy loss.

Schervish. I am very glad to read that Professor Schervish appreciates the importance of invariant procedures. In teaching, I often start my lectures by stating that any inferential

procedure which is not invariant under monotonic transformations of the parameter is suspect, and go on to provide a set of examples of those as "counterexamples" to common statistical procedures.

I agree with Professor Schervish on the importance of the likelihood principle, but I believe that the principle is actually compatible with a sensible use of reference distributions. Indeed, a reference posterior encapsulates, by definition, the (minimal) inferential statements you could proclaim about the parameter of a model *if* your prior was that maximizing the information that data generated from that *particular* model could possibly provide. If you change the model (even if the new model induces a proportional likelihood function), you change the reference prior. Thus, different reference posteriors corresponding to different sampling schemes with Bernoulli observations provide a collection of *conditional* answers (one for each sampling scheme one is willing to consider), which may all be part of the sensitivity analysis to changes in the prior mentioned above.

Objectivity is indeed an emotionally charged word, and it should be explicitly qualified whenever it is used. No statistical analysis is seriously objective, if only because the choice of both the experiment design and the model used have typically very strong subjective inputs. However, the frequentist paradigm is sold as "objective" just because its conclusions are only conditional on the model assumed and the data obtained, and this objectivity illusion has historically helped frequentist to keep a large share of the statistics market. I claim for the procedures described in this paper the right to use "objective" in precisely the same sense: these are procedures which are only conditional on the assumed model and the observed data. The use of the word "objective" in this precise, limited sense may benefit, I believe, the propagation of the Bayesian paradigm. For a recent discussion of this and related issues see Berger (2006) and ensuing discussion.

I fully agree with Professor Schervish on the paramount importance of clearly presenting the assumptions needed for an inferential statement. In the case of reference posteriors this should typically read as a *conditional* statement of the form: "*If* available data $\boldsymbol{x}$ had been generated by model $\mathcal{M} \equiv \{p_x(\cdot \mid \boldsymbol{\omega}),\ \boldsymbol{\omega} \in \Omega\}$ *and* prior information about $\boldsymbol{\theta}(\boldsymbol{\omega})$ were minimal with respect to the information about $\boldsymbol{\theta}(\boldsymbol{\omega})$ that repeated sampling from $\mathcal{M}$ could possibly provide *then*, the marginal reference posterior $\pi(\boldsymbol{\theta} \mid \boldsymbol{x})$ encapsulates what could be said about the value of $\boldsymbol{\theta}$, solely on the basis of that information".

References

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1, 385-402 and 457-464 (with discussion).